Doctoral Thesis in agreement with the cotutelle contract between

Université Pierre et Marie Curie

Doctoral school: Complexité du vivant - ED515
Laboratory: Center of Research in Myology – Institute of Myology

and

Freie Universität Berlin

Department of Biology, Chemistry and Pharmacy
Laboratory: Max Delbrück Center for Molecular Medicine

Bioinformatics tools for the systems biology of dysferlin deficiency

Apostolos Malatras

Supervised by Dr. Gillian Butler-Browne, Prof. Simone Spuler and Dr. William Duddy

Publicly presented and defended on 13th December 2017

PhD Jury composition:

Dr. Gillian Butler-Browne (Thesis co-supervisor)

Prof. Dr. med. Simone Spuler (Thesis co-supervisor)

Dr. William Duddy (Thesis co-supervisor)

Dr. George Paliouras (Reviewer)

Prof. Yetrib Hathout (Reviewer)

Prof. Frédéric Devaux (UPMC representative)

Prof. Dr. Sigmar Stricker (FU representative)

Dr. Maria Reichenbach (FU Postdoc)

Bioinformatics tools for the systems biology of dysferlin deficiency

Dissertation

In fulfillment of the requirements for the degree
"Doctor rerum naturalium (Dr. rer. nat.)"
integrated in the International Graduate School for Myology MyoGrad
in the Department for Biology, Chemistry and Pharmacy at the Freie Universität Berlin
in Cotutelle Agreement with the Ecole Doctorale 515 "Complexité du Vivant" at the
Université Pierre et Marie Curie Paris

Submitted by Apostolos Malatras
Berlin, 2017

Supervisor: Prof. Dr. med. Simone Spuler
Second examiner: Prof. Dr. Sigmar Stricker
Date of defense: 13/12/2017

Dedicated to

Anna, Giorgos and Giota

Acknowledgments

First of all, my thanks go to my parents for their constant support and love. They taught me valuable life lessons on how to be strong, patient and confident but also gave me the opportunity to follow my dreams. I especially thank my sister who greatly helped me during my studies.

I am especially grateful to my supervisor William Duddy for entrusting me with this project and also giving me the opportunity to work in a great research environment. His excellent guidance, motivation and advice, throughout my PhD studies, contributed enormously to this work. His brilliant thinking and careful comments made me think harder and out of the box. I am also grateful for granting me the freedom to approach computational problems using multiple methods.

Particular thanks go to the director of my thesis in Paris, Gillian Butler-Browne, for her valuable guidance and support throughout this work. Special thanks go to Simone Spuler, the director of my thesis in Berlin, for her support and trust in my work.

Furthermore, I thank the members of the 1st and 2nd PhD committees: Thierry Jaffredo, David Salgado, and Philippe Chavrier for their helpful comments. I am grateful for the reviewers of my thesis manuscript, George Paliouras and Yetrib Hathout for their careful comments and constructive criticism. Special thanks to Sigmar Stricker, Frédéric Davaux and Maria Reichenbach who accepted to be part of the thesis jury. I am also very grateful to Ioannis Michalopoulos for having constructive talks on data analysis techniques.

I would like to thank my office colleagues in Paris: Gonzalo Cordova, Pierre Klein, Jessy Etienne and Eliza Negroni who transformed the office in a relaxed and fun workplace.

I would also like to thank the members of the Spuler group in Berlin that made my stay there enjoyable. Special thanks to Stefanie Grunwald who shared her office with me and made the transition from Paris to Berlin very easy.

I thank the MyoGrad research program for creating this unique collaboration and offering research experience between the Center of Research in Myology in Paris and the Experimental and Clinical Research Center in Berlin. I thank Heike Pascal, Gisele Bonne, Lidia Dolle and Soraya Sandal for helping with the administrative work in Paris. My special thanks goes to Susanne Wissler who helped immensely with the cotutelle and thesis complex processes, and also being patient and understanding on every administrative request I had.

Finally, my stay in Paris would never have been enjoyable without the support of my friends and colleagues. I would like to thank all the people who shared a little bit of their lives with me, while helping me understand myself better. In particular, I thank Matthew Thorley, Coline Macquart, Maria Chatzifrangkeskou and Nada Essawy for the countless hours we spent in or outside of the lab, the holidays we have had together, eating healthy at crous and having fun at parties. I will definitely miss your company and really hope to work alongside the breakthrough team in the future. I would also like to thank Teresa Gerhalter, Blanca Rodriguez Morales, Daniel Owens, Magdalena Matłoka, Damily de Dea Diniz and Margot Saunier for all the good times we had together.

Abstract

The aim of this project was to build and apply tools for the analysis of muscle omics data, with a focus on Dysferlin deficiency. This protein is expressed mainly in skeletal and cardiac muscles, and its loss due to mutation (autosomal-recessive) of the *DYSF* gene, results in a progressive muscular dystrophy (Limb Girdle Muscular Dystrophy type 2B (LGMD2B), Miyoshi myopathy and distal myopathy with tibialis anterior onset (DMAT)). We have developed various tools and pipelines that can be applied towards a bioinformatics functional analysis of omics data in muscular dystrophies and neuromuscular disorders. These include: tests for enrichment of gene sets derived from previously published muscle microarray data and networking analysis of functional associations between altered transcripts/proteins. To accomplish this, we analyzed hundreds of published omics data from public repositories. The tools we developed are called CellWhere and MyoMiner.

CellWhere is a user-friendly tool that combines protein-protein interactions and protein subcellular localizations on an interactive graphical display. It accepts a list of genes and generates a protein-protein interaction network graph organized into subcellular locations to mimic the structure of the cell. Localization annotations acquired from the manually curated public repositories, Gene Ontology and UniProt (Swissprot), are mapped to a smaller number of CellWhere localizations. Protein-protein interactions and their scores are acquired from the Mentha interactome server. CellWhere can be accessed freely at https://cellwhere-myo.rhcloud.com

MyoMiner is a muscle cell- and tissue-specific database that provides co-expression analyses in both normal and pathological tissues. Many gene co-expression databases already exist and are used broadly by researchers, but MyoMiner is the first muscle-specific tool of its kind. High-throughput microarray experiments measure mRNA levels for thousands of genes in a biological sample and most microarray studies are focused on differentially expressed genes. Another way of using microarray data is to exploit gene co-expression, which is widely used to study gene regulation and

function, protein interactions and signaling pathways. These co-expression analyses will help muscle researchers to delineate muscle pathology specific protein interactions and pathways. Changes in co-expression between pathologic and healthy tissue may suggest new disease mechanisms and therapeutic targets. MyoMiner is a powerful muscle specific database for the discovery of genes that are associated in related functions based on their co-expression and is available at https://myominer-myo.rhcloud.com.

These tools will be used in the analysis and interpretation of transcriptomics data from dysferlinopathic muscle and other neuromuscular conditions and will be important to understand the molecular mechanisms underlying these pathologies.

Résumé

Le but de mon projet est de créer et d'appliquer des outils pour l'analyse de la biologie des systèmes musculaires en utilisant différentes données OMICS. Ce projet s'intéresse plus particulièrement à la dysferlinopathie due la déficience d'une protéine appelée dysferline qui est exprimée principalement dans les muscles squelettiques et cardiaque. La perte du dysferline due à la mutation (autosomique-récessive) du gène *DYSF* entraîne une dystrophie musculaire progressive (LGMD2B, myopathie Miyoshi, DMAT).

Nous avons déjà développé des outils bio-informatiques qui peuvent être utilisés pour l'analyse fonctionnelle de données OMICs, relative à la dyspherlinopathie. Ces derniers incluent le test dit «gene set enrichment analysis», test comparant les profils OMICS d'intérêts aux données OMICS musculaires préalablement publiées ; et l'analyse des réseaux impliquant les diffèrent(e)s protéines et transcrits entre eux/elles. Ainsi, nous avons analysé des centaines de données omiques publiées provenant d'archives publiques. Les outils informatiques que nous avons développés sont CellWhere et MyoMiner.

CellWhere est un outil facile à utiliser, permettant de visualiser sur un graphe intéractif à la fois les interactions protéine-protéine et la localisation subcellulaire des protéines. En résumé, après avoir téléchargé une liste de gènes d'intérêts, CellWhere génère des graphes de réseaux d'interaction entre protéines. Ces réseaux sont alors représentés dans les différents compartiments subcellulaires, mimant ainsi la structure de la cellule. Les localisations subcellulaires détaillées sont obtenues à partir de banques de données telles que Gene Ontology et UniProt, puis sont regroupées en compartiments subcellulaires sur CellWhere, permettant ainsi une meilleure lisibilité des graphes. Les interactions proteines-proteines et leurs scores sont obtenus à partir du serveur d'interactomes Mentha. Il est possible d'accéder à CellWhere via ce lien : https://cellwhere-myo.rhcloud.com

Myominer est une base de données spécialisée dans le tissu et les cellules musculaires, et qui fournit une analyse de co-expression, aussi bien dans les tissus sains que pathologiques. Plusieurs bases de données de co-expressions géniques existent déjà pour tous les tissus, et sont très utilisées par les chercheurs, mais Myominer est le premier outil de ce genre, spécialisé dans le muscle. Des expériences de puces à haut débit permettent de mesurer des niveaux d'ARN messagers pour des milliers de gènes dans un échantillon biologique, et la plupart des études sur puces sont focalisées sur les expressions géniques différentielles. Une autre façon d'utiliser les données de micropuces est d'exploiter la co-expression génique, qui est largement utilisée pour étudier la régulation et la fonction des gènes, les interactions protéiques, ainsi que les voies de signalisation. Il est possible d'accéder à Myominer via ce lien : https://myominer-myo.rhcloud.com

Ces outils seront utilisés dans l'analyse et l'interprétation de données transcriptomiques pour les dyspherlinopathies mais également les autres pathologies neuromusculaires. Par ailleurs, ils faciliteront la compréhension des mécanismes moléculaires caractérisants ces maladies.

Zusammenfassung

Das Ziel dieser Arbeit war es Anwendungen für die Systembiologieanalyse von Muskel Omics Daten mit einen Fokus auf Dysferlinopathien zu entwickeln und anzuwenden. Das Dysferlinprotein wird hauptsachlich in der Skeletmuskulatur und im Herzmuskel exprimiert, wobei das Fehlen dieses Proteins, was durch Genmutationen (autosomal rezessiv) im Dysferlingen hervorgerufen wird, zu einer progressiven Muskeldystophie (LGMD 2B, Myoshi Myopathie, DMAT) führt. Wir entwickelten verschiedene Tools und Pipelines, die für die bioinformatische Funktionalanalyse von Omics Daten in Dysferlinopathien und neuromuskuläre Erkrankungen verwendet werden können. Unter anderem, einen Test für Anreicherungen von Gensets, welche von früher publizierten Muskelarraydaten stammen, und eine Netzwerkanalyse für funktionelle Assoziierung zwischen veränderten Transkripten und Proteine. Für die Realisation dieser Projekte analysierten wir hunderte von publizierten Omics Dateien von öffentlich zugänglichen Dateibanken und entwickelten die Tools CellWhere und MyoMiner.

CellWhere ist ein anwenderfreundliches Tool, welches die Protein-Protein Interaktionen und die subzelluläre Lokalisierung der Proteine in einer interaktiven Grafikanzeige darstellt. Es verwendet eine Liste an Genen und generiert eine Protein-Protein Interaktionsnetzwerkgraphik, welche dann deren subzelluläre Lokalisierung darstellt während es die Zellstruktur imitiert. Lokalisierungsannotationen werden von den gemeinschaftlich gewarteten Datenbanken, Gene Ontology und Uniprot, bezogen und an eine kleinere Anzahl von CellWhere Lokationen zugeordnet. Die Protein-Protein Interaktionen und deren Werte stammen vom Mentha interactome Server. CellWhere ist frei zugänglich unter: https://cellwhere-myo.rhcloud.com

MyoMiner ist eine Muskelzell und –gewebe spezifische Datenbank, die eine Expressionsanalyse von gesunden und pathologischen Gewebe anbietet. Viele Gen-Co-Expression Datenbanken sind heutzutage zugänglich und werden auch umfassend von Wissenschaftlern genutzt. MyoMiner ist jedoch das erste muskelspezifische Tool seiner

Art. Hochdurchsatz Microarray Experimente messen die mRNA Levels von tausenden Gene in einer biologischen Probe, wobei die meisten Microarraystudien auf die differentielle Genexpression ausgerichtet sind. Eine andere Möglichkeit diese Microarraydaten zu verwenden ist die Untersuchung von Gen-Co-Expression, welche oft genutzt wird um Genregulierung und –funktion, Proteininteraktionen und Signalketten zu erforschen. Solche Co-Expressions-Analysen werden den Muskelforschern helfen die gewebs-, zell- und pathologiespezifischen Elemente der Muskelproteininteraktionen, Zellsignalen und Genregulierungen zu beschreiben. Änderungen in der Co-Expression zwischen kranken und gesunden Gewebe könnten dann im Weiteren neue Krankheitsmechanismen und Therapieansätze vorbringen. MyoMiner ist eine mächtige muskelspezifische Datenbank; ausgelegt für die Erforschung von Genen, die in verwandten Funktionen aufgrund ihrer Co-Expression verbunden sind. Diese Datenbank steht zur Verfügung unter: https://myominer-myo.rhcloud.com

Diese Tools werden für die Analyse und die Interpretation von Transkriptom Daten vom Dysferlin-defizienten Muskelgewebe und anderen neuromuskulären Erkrankungen verwendet und sind wichtig, um die molekularen Mechanismen der zugrundeliegenden Pathologien zu verstehen.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

AAV	Adeno-associated viral vector	Limma	Linear Models for Microarray and RNA-seq Data
AE	Array express	MAS 5.0	Affymetrix microarray suite 5
ANXA1	Annexin A1	MD	Muscular dystrophy
ANXA2	Annexin A2	MDS	multidimensional scaling
arctanh	Inverse hyperbolic tangent function	MI	Mutual information
BH	Benjamini-Hochberg	MIAME	Minimum information about a microarray experiment
BMD	Becker muscular dystrophy	MM	Miyoshi myopathy, Mismatch probe
CAPN3	Calpain-3	MMD1	Miyoshi myopathy
CAV3	Caveolin-3	*MYOF*	Myoferlin
CDF	Chip description file	NGS	Next generation Sequencing
cDNA	complementary DNA	NUSE	Normalized unscaled standard error
CEL	Affymetrix microarray intensity files, also referred as raw files	*OTOF*	Otoferlin
CK	Creatine kinase	*PARVB*	Beta-parvin or affixin
DGC	Dystrophin-associated glycoprotein complex	PCA	Principal components analysis
DM2	Diabetes mellitus type 2	PDB	Protein data bank
DMAT	Distal myopathy with anterior tibialis onset	PLIER	Probe logarithmic intensity error
DMD	Gene symbol for dystrophin but also abbreviation for Duchenne muscular dystrophy,	PM	Perfect match probe
DYSF	Gene symbol for dysferlin	PPI	Protein-protein interactions
EBI	European bioinformatics institute	QC	Quality controls
FDR	False discovery rate	RLE	Relative log expression
FSHD	Facioscapulohumeral muscular dystrophy	RMA	Robust multi-array average
GC-RMA	GeneChip RMA	SAGE	Serial analysis of gene expression
GDS	GEO datasets	SAM	Significance analysis of microarrays
GEO	Gene expression omnibus	SCAN	Single channel array normalization
GO	Gene ontology	SDRF	Sample and data relationship format

GPL	GEO platforms	SNP	Single nucleotide polymorphisms
GPL1261	GEO ID for Affymetrix MG 430 2.0 array	SOFT	Simple omnibus format in text
GPL570	GEO ID for Affymetrix HG-U133 Plus 2.0 array	*SYNPO*	Synaptopodin
GSE	GEO series	*SYNPO2*	Synaptopodin-2
GSM	GEO samples	*SYNPO2L*	Synaptopodin-2 like protein
GTEx	Genotype tissue expression project	*SYT*	Synaptotagmins
IM	Ideal mismatch	TF	Transcription factor
KEGG	Kyoto encyclopedia of genes and genomes	tahn	hyperbolic tangent function
k-NN	k-nearest neighbor classifier	TRIM	Tripartite motif protein family
KO	knockout	UML	Unified modeling language
LGMD	Limb girdle muscular dystrophy	UPC	Universal expression code

Preamble

Since the explosion of high-throughput technologies, a huge collection of data is available for researchers, but the processing of this and extraction of information from it remains a major challenge. In this work, we set out to retrieve and combine muscle-specific raw data from public repositories, assess their quality and develop a robust analysis pipeline, which will give consistent and comparable results and systems biology tools for muscle research. We also set out to analyze the acquired information in the context of Dysferlin deficiency. This collection of muscle data can complement our functional understanding of muscle-specific genes, suggest networks of biological interactions, and enable us to identify sets of genes that are regulated in different conditions of muscle and muscle neuromuscular pathology.

This doctoral thesis comprises four parts:

a) A general background on muscular dystrophies with emphasis on dysferlinopathies, and an introduction to the field of omics with emphasis on microarray transcriptomics technology.

b) Three manuscripts: two peer-reviewed articles and one in preparation, which document and discuss the scientific work that has been conducted during this PhD. Two of the manuscripts describe systems biology tools specific to the field of myology: CellWhere and MyoMiner.

c) A discussion chapter that summarizes the principal outcomes of the thesis.

d) An appendix with two additional peer-reviewed manuscripts: a review on muscle aging, and a crowdsourced article for the extraction of gene signatures across multiple microarray samples.

Chapter 1 – Introduction

1.1.1 Muscular dystrophies

Muscular dystrophies (MD) are a diverse group of inherited diseases that share features of progressive weakness and wasting of the muscle tissue (Table 1). Although MDs are known for the selective involvement of skeletal muscles, other abnormalities can be detected in various tissues such as the cardiac muscles, the respiratory system, smooth muscles, neurons and the brain (Coral-Vazquez, Cohn et al. 1999; Moore, Saito et al. 2002). They have traditionally been classified according to the clinical findings, inheritance type, onset age of the disease, affected muscle group, and overall progression (Cohn and Campbell 2000). Heterogeneous groups such as congenital and limb girdle muscular dystrophies (LGMDs) were classified to different subtypes based on inheritance and genetic defects. Better understanding of the mechanisms involved in MDs gave insights that the classification cannot be based only on the aforementioned methods, since some phenotypes are associated with mutations in different but functionally similar genes (Guglieri, Straub et al. 2008; Mercuri and Muntoni 2012).

After the discovery of the dystrophin gene (*DMD*) (Hoffman, Brown et al. 1987), many more genes were identified as being linked to various muscular dystrophies. Most of the common MDs are related to genes that encode components of the Dystrophin-associated Glycoprotein Complex (DGC) which links the intracellular actin cytoskeleton with the extracellular matrix. A defect in a protein belonging to this complex can destabilize the whole complex which makes the muscle cell membrane (sarcolemma) susceptible to contraction injuries which in turn leads to muscle necrosis (Petrof, Shrager et al. 1993). This shows that it is very important to maintain the plasma membrane structural integrity of muscle cells in order to have normal function. Thus, mechanisms to repair the sarcolemma (mend the physical injuries) were evolved (McNeil and Steinhardt 1997; Meldolesi 2003).

Other forms of muscular dystrophy arise from mutations in genes that are unrelated to the DGC. For example, defective plasma membrane repair can cause muscle wasting and will lead to muscular dystrophy. Dysferlin gene (*DYSF*), even though it expresses dysferlin protein that it is not a part of the DGC complex (Bansal, Miyake et al. 2003), was identified as the mutant gene that caused clinically distinct muscular dystrophies called dysferlinopathies also known as limb girdle musclular dystrophy (LGMD), miyoshi myopathy (MM) and distal myopathy with anterior tibialis onset (DMAT) (Bashir, Britton et al. 1998; Liu, Aoki et al. 1998; Illa, Serrano-Munuera et al. 2001). Studies show that Ca^{2+}-dependent membrane repair in skeletal muscles is disrupted with loss of dysferlin that results in a slowly progressive muscle weakness and necrosis (Bansal, Miyake et al. 2003; Lennon, Kho et al. 2003).

The most common MD with childhood onset is Duchenne muscular dystrophy (DMD) and its milder form, Becker muscular dystrophy (BMD), both affecting about 1 out of 5,000 boys. The most common MD with adult age of onset is Myotonic dystrophy which affects about 1 per 10,000 men followed by facioscapulohumeral muscular dystrophy (FSHD) that affects about 3 in 100,000 men. The recessive forms of Limb girdle muscular dystrophies (LGMD) are much more common than the dominant ones with a 9 to 1 ratio (Thompson and Straub 2016). The frequencies of certain muscular dystrophies vary by region. For example, LGMD2A is more common in southern Europe (Fanin, Nascimbeni et al. 2005), while LGMD2I and 2B are common in northern Europe (Sveen, Schwartz et al. 2006).

Type	Description
Duchenne muscular dystrophy (DMD)	The most common childhood onset (2 to 6 years old) muscular dystrophy. Mutation or loss of dystrophin gene causes Duchenne muscular dystrophy. The gene is located on the X chromosome thus affecting only boys (with rare exceptions). Symptoms include muscle wasting, necrosis and weakness. Affected muscles are lower, upper limbs and pelvis which eventually spread to all skeletal muscles. The disease progresses rapidly. Over the past decades, survival of DMD

	patients has improved to mid 20s and 30s (Passamano, Taglia et al. 2012).
Becker muscular dystrophy (BMD)	Almost identical to Duchenne muscular dystrophy with less severe symptoms. Slower progress than Duchenne. Lifespan range from late adulthood to old age (Lovering, Porter et al. 2005).
Congenital muscular dystrophy (CMD)	Starts at birth. Multiple organs are involved including the brain. Muscle weakness could be mild or severe affecting all voluntary muscles. The disease progress is slow. Patient's lifespan is generally shortened (Mercuri and Muntoni 2012).
Emery-Dreifuss muscular dystrophy (EDMD)	Age of onset is typically on childhood to early adulthood. Divided into three subtypes: X-linked, autosomal, dominant and recessive with the first one being the most common. Emery-Dreifuss MD is caused by mutations in the LMNA or EMD gene. Symptoms include muscle weakness and wasting of upper limbs and shin muscles. Disease progress is slow, but due to problems in normal cardiac function sudden death may occur (Ostlund and Worman 2003).
Facioscapulohumeral muscular dystrophy (FSHD)	Usually starts from childhood to adulthood. Affected muscles are: facial, shoulders, and upper limbs. The progress is slow (rapid deterioration periods are possible). The patients usually live to old age (Lemmers, Wohlgemuth et al. 2007).
Limb-Girdle muscular dystrophy (LGMD)	Age of onset is childhood to adulthood. The first affected muscles are the shoulder and pelvic girdle. More than 30 subtypes of LGMDs have been reported. They are classified based on inheritance in autosomal dominant and autosomal recessive. The latter is more common with more severe symptoms. The disease's progress is slow with patients living into old age, even though being non-ambulatory. Usual cause of death is cardiopulmonary problems (Nigro and Savarese 2014).
Distal muscular dystrophy (DD)	Age of onset varies from early adulthood to old age. The affected muscles are the lower legs (calf) and the forearms. Two of the Distal muscular dystrophies (Miyoshi myopathy and distal myopathy with tibial anterior onset) are caused by loss of dysferlin which is also responsible for the type 2B limb girdle muscular dystrophy. Disease progress is slow and rarely is lethal (Udd 2011).

Myotonic muscular dystrophy (MMD)	It appears on adults. It is an autosomal dominant disease that affects the face, neck, and foot muscle groups first and then spreads to all muscles. Clinical characteristics are muscle wasting and weakness alongside delayed relaxation of muscles after contraction (myotonia). The disease progress is slow with patients living to old age (Turner and Hilton-Jones 2010).
Oculopharyngeal muscular dystrophy (OPMD)	Age of onset is typically late adulthood. Symptoms include weakness and degeneration of eyelid, face and throat muscles first and then shoulder and pelvic girdle muscles. Disease progress is slow (Trollet, Gidaro et al. 1993).

Table 1 | Muscular dystrophy types. Muscular dystrophies are classified into 9 types according to their characteristics.

1.1.2 Limb-Girdle Muscular Dystrophies (LGMDs)

LGMDs (Walton and Nattrass 1954) are a group of phenotypically and genotypically heterogeneous rare muscular dystrophies. They are typically characterized by predominant atrophy and weakness of the proximal muscles (shoulders and pelvic girdle) of the lower and upper limbs. Cardiac, respiratory and other muscles are often affected (Verhaert, Richards et al. 2011). The age of onset is usually between childhood and early adulthood, but for some patients the disease begins much later. If the myopathy starts in childhood it progresses rapidly, with a more severe and disabling form. If it begins on adulthood, it progresses more slowly with milder symptoms allowing some patients to have a fairly normal life (Nigro, Aurino et al. 2011).

LGMDs are separated into two groups: autosomal dominant inheritance where a mutant gene from one parent is sufficient to cause the disease, called type 1 LGMD (LGMD1) and the autosomal recessive inheritance where defects or mutations on both alleles (both parents) are required, called type 2 LGMD (LGMD2). Type 1 LGMDs typically begin at early adulthood, often exhibit a mild phenotype and are considered rare as they represent about 10% of all LGMDs. The vast majority of LGMDs are of type 2 with more severe symptoms and disease course. At the time of writing this thesis, more than 30 different LGMD subtypes have been discovered (Table 2) (Nigro and

Savarese 2014; Thompson and Straub 2016). All have distinct genetics and a wide variety of phenotypes.

Mutation in the skeletal muscle sarcoglycan complex genes, *SGCG-A-B-C*, cause LGMD2C-D-E-F, respectively (Table 2). Sarcoglycans, members of the dystrophin-complex, are n-glycosyl transmembrane proteins with a large extracellular, a transmembrane, and a short intracellular domain. These disorders, also called sarcoglycanopathies, have some similarities with Duchenne and Becker dystrophies such as early childhood disease onset with both heart and respiratory muscles being affected. Another subgroup is the dystroglycanopathies LGMD2I-J-M-N-O-P with mutations to their respective genes: *POMT1*, *POMT2*, *POMGNT1*, *FKTN*, *FKRP* and *DAG1* (Table 2) (Muntoni, Torelli et al. 2011). LGMD2B is caused by mutation in the DYSF gene (dysferlin) (Liu, Aoki et al. 1998). The diseases related to dysferlin mutations are also called dysferlinopathies and include LGMD2B (proximal onset), Miyoshi myopathy (MM) (distal onset), distal myopathy with anterior tibialis onset (DMAT), and other phenotypes. However, they are not classified based on different mutations of dysferlin. Patients have normal mobility in childhood as the onset is usually in early adulthood, although as with the other LGMDs the symptoms range from severe (childhood onset) to mild (late onset) (Urtizberea, Bassez et al. 2008).

LGMDs can be diagnosed via a combination of a broad range of procedures and tests that include: clinical assessment, electromyography, muscle biopsy that shows dystrophic changes indicative of de- and re-generation of muscle fibers, very high creatine kinase (CK) levels due to myofibre damage and necrosis, genetic testing, and immunohistochemical tests to determine the absence of the protein involved and thus the type of muscular dystrophy (Laval and Bushby 2004; Narayanaswami, Carter et al. 2015).

Type	OMIM ID	Gene	Reference
LGMD1A	159000	*TTID*	(Hauser, Horrigan et al. 2000)
LGMD1B	159001	*LMNA*	(Muchir, Bonne et al. 2000)
LGMD1C	607801	*CAV3*	(McNally, de Sa Moreira et al. 1998; Minetti,

			Sotgia et al. 1998)
LGMD1D	603511	*DNAJB6*	(Harms, Sommerville et al. 2012; Sarparanta, Jonson et al. 2012)
LGMD1E	601419	*DES*	(Greenberg, Salajegheh et al. 2012; Hedberg, Melberg et al. 2012)
LGMD1F	608423	*TNPO3*	(Melia, Kubota et al. 2013; Torella, Fanin et al. 2013)
LGMD1G	609115	*HNRPDL*	(Vieira, Naslavsky et al. 2014)
LGMD1H	613530	*3p23-p25*	(Bisceglia, Zoccolella et al. 2010)
LGMD2A	253600	*CAPN3*	(Richard, Broux et al. 1995)
LGMD2B	253601	*DYSF*	(Bashir, Britton et al. 1998; Liu, Aoki et al. 1998)
LGMD2C	253700	*SGCG*	(Noguchi, McNally et al. 1995)
LGMD2D	608099	*SGCA*	(Roberds, Leturcq et al. 1994)
LGMD2E	604286	*SGCB*	(Lim, Duclos et al. 1995)
LGMD2F	601287	*SGCD*	(Nigro, de Sa Moreira et al. 1996)
LGMD2G	601954	*TCAP*	(Moreira, Wiltshire et al. 2000)
LGMD2H	254110	*TRIM32*	(Frosk, Weiler et al. 2002)
LGMD2I	607155	*FKRP*	(Brockington, Blake et al. 2001)
LGMD2J	608807	*TTN*	(Hackman, Vihola et al. 2002)
LGMD2K	609308	*POMT1*	(Balci, Uyanik et al. 2005)
LGMD2L	611307	*ANO5*	(Bolduc, Marlow et al. 2010)
LGMD2M	611588	*FKTN*	(Godfrey, Escolar et al. 2006)
LGMD2N	607439	*POMT2*	(Biancheri, Falace et al. 2007)
LGMD2O	606822	*POMGNT1*	(Clement, Godfrey et al. 2008)
LGMD2P	613817	*DAG1*	(Hara, Balci-Hayta et al. 2011)
LGMD2Q	613723	*PLEC1*	(Gundesli, Talim et al. 2010)
LGMD2R	615325	*DES*	(Cetin, Balci-Hayta et al. 2013)
LGMD2S	615356	*TRAPPC11*	(Bogershausen, Shahrzad et al. 2013)
LGMD2T	615352	*GMPPB*	(Carss, Stevens et al. 2013)
LGMD2U	616052	*ISPC*	(Tasca, Moro et al. 2013)
LGMD2V	NA	*GAA*	(Preisler, Lukacs et al. 2013)
LGMD2W	616827	*LIMS2*	(Chardon, Smith et al. 2015)
LGMD2X	616812	*BVES*	(Schindler, Scotton et al. 2016)

Table 2 | Limb Girdle Muscular Dystrophy family. The LGMD1 type are autosomal dominant and less severe than the LGMD2 which are autosomal recessive.

1.1.3 Dysferlin

Using a positional cloning strategy, Dysferlin was identified as the gene involved in LGMD 2B and Miyoshi Myopathy muscular dystrophies (Bashir, Britton et al. 1998; Liu, Aoki et al. 1998). The human Dysferlin gene (*DYSF*, *FER1L1*) is at chromosomal location 2p13.2. Dysferlin's original (canonical) isoform consists of 55 exons ranging in size from 30 to 365 base pairs, comprising a total of 6796 bp in length (Human assembly GRCh38.p10, Ensembl transcript DYSF-201; ENST00000258104.7 (Aken, Ayling et al. 2016)). Intron lengths range from less than 200 to more than 30,000 bp. The total length of the dysferlin gene is 223,047 bp. It is expressed in many tissues and cells including skeletal muscles, heart, brain, spleen, placenta, myoblasts, myotubes, at lower levels in lung, kidney and liver, and most highly at skeletal, cardiac muscles and whole blood (Figure 1) (Anderson, Davison et al. 1999; Klinge, Laval et al. 2007).

Dysferlin is a member of the ferlin-1 like (FER1L or simply ferlin) protein family. Proteins that belong to the FER1L family show structural similarities and sequence homology with the *C. elegans* fer-1 protein, which is mainly expressed in spermatocytes. Defects in fer-1 prevent spermatic vesicle fusion, resulting in infertile sperm (Achanzar and Ward 1997). Dysferlin is a 2080 amino acids single-pass type II transmembrane protein with a 237,295 Da mass, making it one of the largest human proteins (Uniprot AC: O75923, ID: DYSF_HUMAN) (Liu, Aoki et al. 1998). It contains seven highly conserved C2 domains (C2A-G) which reside in the cytoplasm, a C-terminal helical transmembrane domain and a C-terminal extracellular domain (Figure 2). Each C2 domain is conserved amongst the rest of the ferlin family protein's corresponding C2 positions, suggesting that each C2 domain has a specific role (Washington and Ward 2006). A single mutation in any of the five C2 domains (A, B, D, E, and G) can lead to muscular dystrophy (Therrien, Dodig et al. 2006). Dysferlin also includes two DysF domains, where one is nested within the other (DysFN and DysFC) and two Fer domains (FerA and FerB) with unknown function (Figure 2). The structure of the first C2 domain of dysferlin (C2A) has been solved (Figure 3). Fuson *et al.* (Fuson, Rice et al. 2014)showed that it changes conformation upon interaction with calcium ions, which is consistent with phospholipids

binding in a Ca^{2+}-dependent manner (Therrien, Di Fulvio et al. 2009) and with dysferlin's role in skeletal muscle membrane repair processes (Bansal, Miyake et al. 2003). It was also shown that dysferlin is expressed less in myoblasts and more in mature myotubes, suggesting a role in muscle differentiation (de Luna, Gallardo et al. 2006). Dysferlin is localized at the plasma membrane and t-tubule network of skeletal muscle cells. It co-localizes with AHNAK1, 2 and PARVB at the site of plasma membrane injury after the accumulation of Ca2+ around the disruption site and interacts with ANXA1 and 2 (Ampong, Imamura et al. 2005; Matsuda, Kameyama et al. 2005). It co-localizes with CACNA1S and BIN1 in the t-tubule during muscle differentiation (Klinge, Laval et al. 2007; Klinge, Harris et al. 2010).

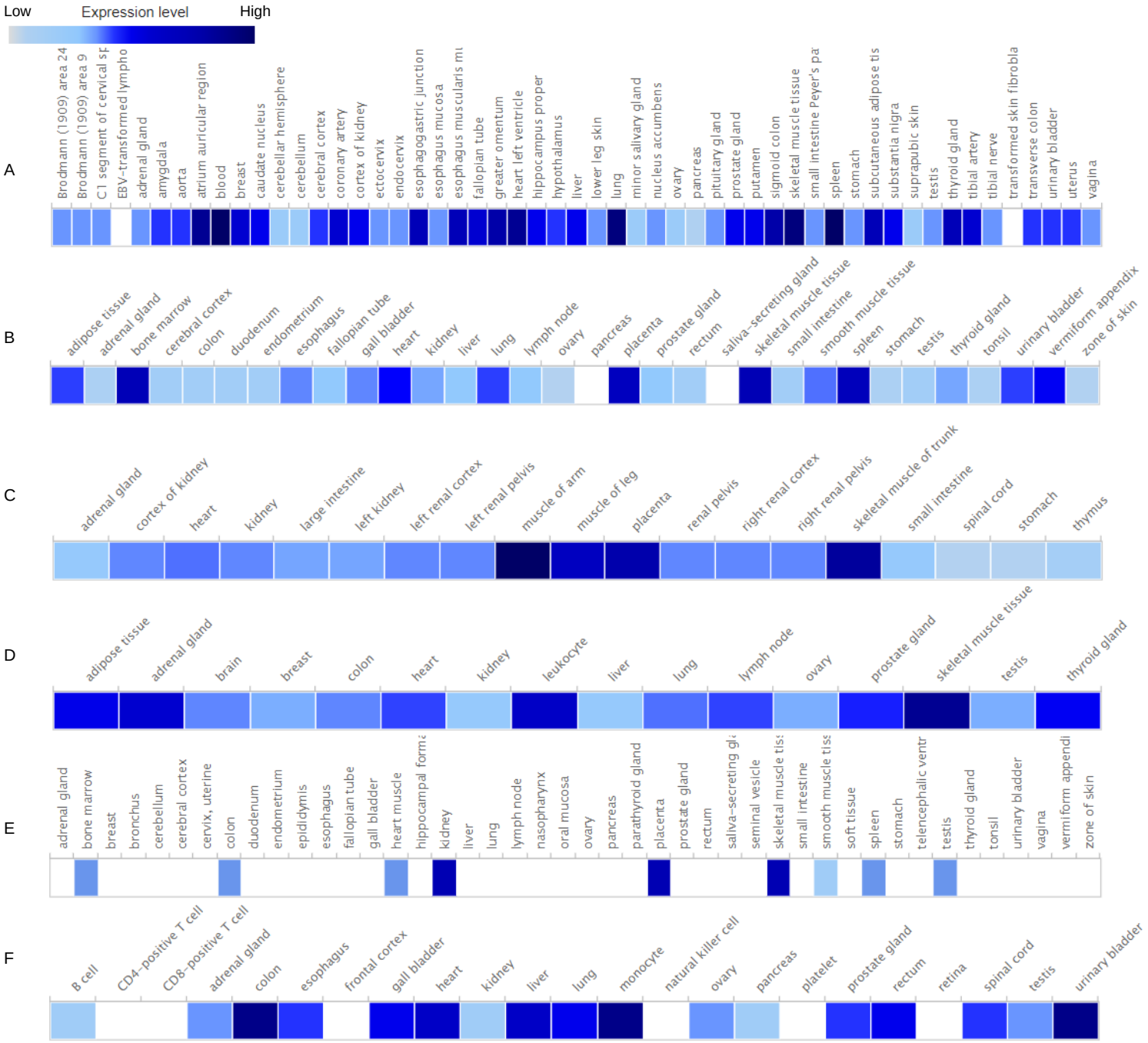

Figure 1 | Gene and protein expression of Dysferlin in various tissues. A) RNA-seq data of 53 human tissue samples from Genotype-Tissue Expression (GTEx) Project (2015). **B)** RNA-seq data from tissue samples of 122 human individuals that represent 32 different tissues (Uhlen, Fagerberg et al. 2015). **C)** RNA-seq data of 19 human tissues from fetuses with congenital defects (Roadmap Epigenomics Consortium) (Kundaje, Meuleman et al. 2015). **D)** RNA-Seq data of human individual tissues and mixture of 16 tissues (Illumina Body Map) (Derrien, Johnson et al. 2012). **E)** Immunochemistry data of 83 different

normal cells from 44 tissues (Human Protein Atlas) (Kim, Pinto et al. 2014). **F)** Mass spectrometry data from the PRIDE project (Kim, Pinto et al. 2014). All data were analyzed in Expression Atlas (Petryszak, Keays et al. 2016).

1.1.4 Ferlin protein family

Following the discovery of dysferlin as the gene responsible for LGMD2B and MM, other genes with similar structure and sequence to dysferlin were reported, and these were classified into a new protein family, the ferlin-1 like proteins. The ferlin-1 like family includes 6 members: FER1L1 or DYSF (dysferlin) (Bashir, Britton et al. 1998; Liu, Aoki et al. 1998), FER1L2 or OTOF (otoferlin) (Yasunaga, Grati et al. 1999; Yasunaga, Grati et al. 2000), FER1L3 or MYOF (myoferlin) (Britton, Freeman et al. 2000; Davis, Delmonte et al. 2000), FER1L4, FER1L5, and FER1L6. Ferlins are separated into two different sub-families based on sequence similarity (Figure 2): the first group includes dysferlin, myoferlin and FER1L5 and the second otoferlin, FER1L4 and FER1L6. All contain highly conserved C2 domains and C-terminal transmembrane helices that are used as anchors to the plasma membrane (Figure 2). Myoferlin and FER1L5 are the proteins that more closely resemble dysferlin. In fact, each C2 domain is more related to the positionally correspondent C2 domain of the other ferlin proteins than to the other C2 domains of the same protein (Washington and Ward 2006). It has been reported that the C2A domain of dysferlin is more than 70% homologous to that of myoferlin but much less homologous (about 15% on average) to the other dysferlin C2 domains (Davis, Doherty et al. 2002). Ferlins and other proteins such as synaptotagmins (SYT), are considered to be involved in vesicle fusion events. Dysferlin is required for repair of the muscle plasma membrane and otoferlin for SNARE-mediated membrane fusion (Beurg, Michalski et al. 2010; Johnson and Chapman 2010). So far, only dysferlin and otoferlin are known to be associated with diseases. The last three ferlin family proteins, FER1L4, FER1L5 and FER1L6, are predicted from human and mouse genomic sequences but have not been described yet.

The spermatogenesis factor FER-1 is expressed only in primary spermatocytes of *Caenorhabditis elegans*. Defects of the FER-1 gene disrupt the fusion of spermatid membranous organelles with the plasma membrane, which results in sterility due to immobilization of the spermatids (Achanzar and Ward 1997). A single mutation in any of the three C2 domains alters the Ca^{2+} sensitivity of FER-1, disrupting the fusion of membranous organelles (Washington and Ward 2006; Han and Campbell 2007).

Otoferlin is smaller than dysferlin and has 64% sequence similarity to dysferlin (Bansal and Campbell 2004). It is expressed in the cochlea, brain and vestibule with low expression levels in kidney, lung, skeletal and cardiac muscles (Yasunaga, Grati et al. 2000). Mutations in otoferlin results in a recessive deafness form, called DFNB9 (OMIM 601071) (Yasunaga, Grati et al. 1999; Yasunaga, Grati et al. 2000). Otoferlin interacts with SNARE proteins in a Ca^{2+}-dependant manner at the synapses of the cochlear hair in order to trigger exocytosis of neurotransmitter. The pathology is caused by a loss of calcium mediated exocytosis without the disruption of the synaptic vesicle structure (Roux, Safieddine et al. 2006).

Myoferlin was named according to its high sequence homology to dysferlin. Like dysferlin, it is expressed highly in skeletal and cardiac muscles and is present on the skeletal muscle plasma membrane (Davis, Delmonte et al. 2000). However, it is found in the nucleus unlike dysferlin. Both proteins have about identical molecular weight (~230kDa), seven C2 domains and similar FerA, B and DysFN, C domains (Figure 2). Even though both dysferlin and myoferlin are so similar, they participate in different events. Myoferlin is required for myoblast fusion during differentiation (Doherty, Cave et al. 2005). Also, myoferlin is not overexpressed to compensate for the lack of dysferlin in dysferlin deficient patients, supporting that both proteins have very few overlapping functions (Inoue, Wakayama et al. 2006). Lack of myoferlin causes muscle atrophy in mice, although myoferlin has not yet been linked to any human diseases.

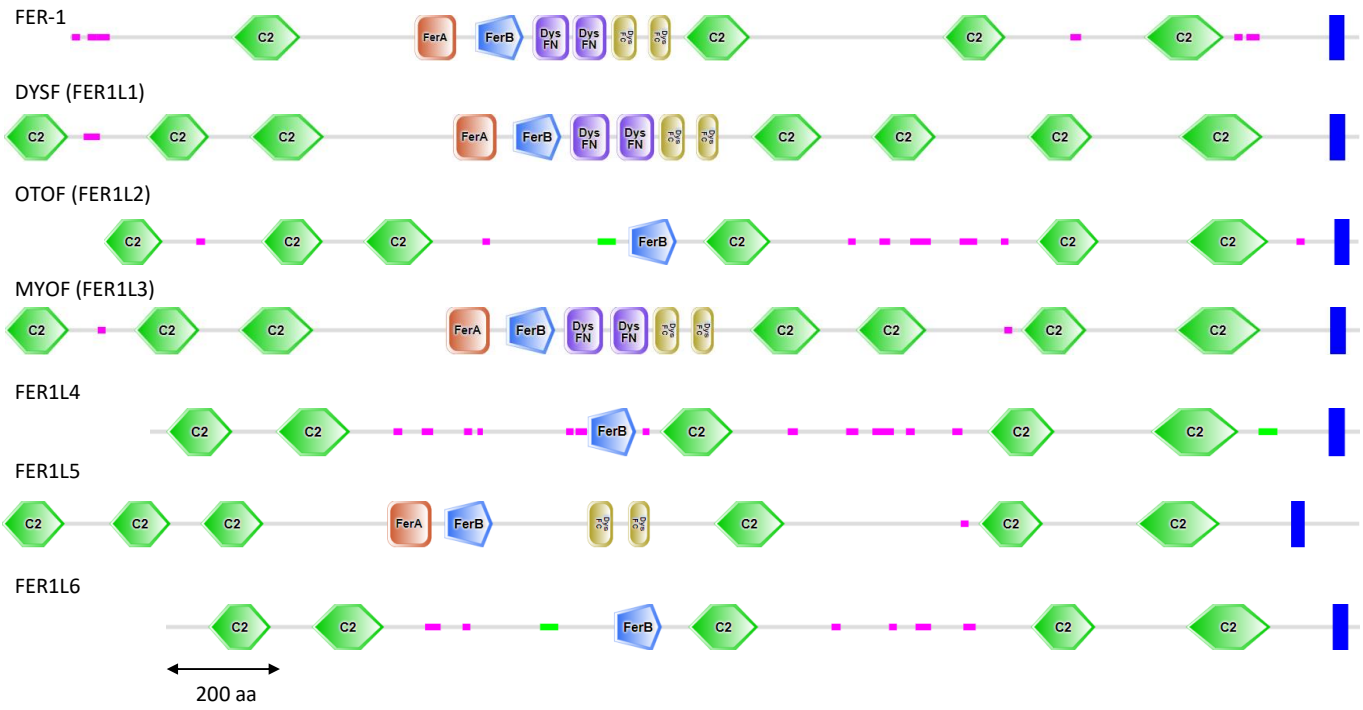

Figure 2 | Ferlin family domain and structural characteristics. *C. elegans* fer-1 protein is on top and the human ferlin-1 like proteins are aligned underneath. The green hexagons represent the conserved C2 domain, FerA is colored red and FerB is blue. DysFN and DysFC are purple and yellow respectively. The C-terminal transmembrane helical region is represented by the blue rectangle. The bright pink rectangles on the grey lines represent low complexity regions and the bright green ones represent coiled coils regions. FER-1, DYSF and MYOF have FerA, FerB and nested DysFN and DysFC domains. FER1L5 includes all the aforementioned domains except for the DysFN. OTOF, FER1L4 and FER1L6 are smaller in length and include only the FerB domain. Domain location, architecture and visualization were created with the SMART tool (Letunic, Doerks et al. 2015).

1.1.5 C2 domain

C2 protein structural domains exhibit many functions such as membrane trafficking and fusion, phospholipid binding and signaling (Pallanck 2003). They are called C2 as they were reported as the second conserved sequence (domain) in protein kinase C (Newton 1995). C2 are independently folded protein domains of between 70-150 amino acid residues that form a beta sandwich structure composed of eight beta strands (Figure 3) (Sutton, Davletov et al. 1995). On one of the beta-sandwich ends we

find the Ca^{2+} binding site, which is mediated through a group of aspartic acid residues (Davis, Doherty et al. 2002).

C2 domains are best studied in the protein family of synaptotagmins which contain two C2 domains. Synaptotagmin's first C2A binds Ca^{2+} and anionic phospholipids (Davletov and Sudhof 1993; Chapman and Jahn 1994), while the second interacts with proteins and binds phospholipids (Fernandez, Arac et al. 2001). Calcium binds to one end of the beta-sandwich that involves aspartic acid residues (Rizo and Sudhof 1998), which are also present in dysferlin, myoferlin and otoferlin. It was suggested that their interactions with phospholipids or other proteins are Ca^{2+}-dependant (Davis, Doherty et al. 2002). Especially the C2A domains of dysferlin and myoferlin demonstrate similar Ca^{2+} binding properties to those of synaptotagmins. Because dysferlin's C2A domain is the furthest away from the plasma membrane it is thought that it may attract vesicles that contain ferlin proteins to the membrane (Davis, Doherty et al. 2002). For dysferlin to function properly, all C2 domains are required, likely mediating interactions with other dysferlin interacting proteins (Klinge, Laval et al. 2007).

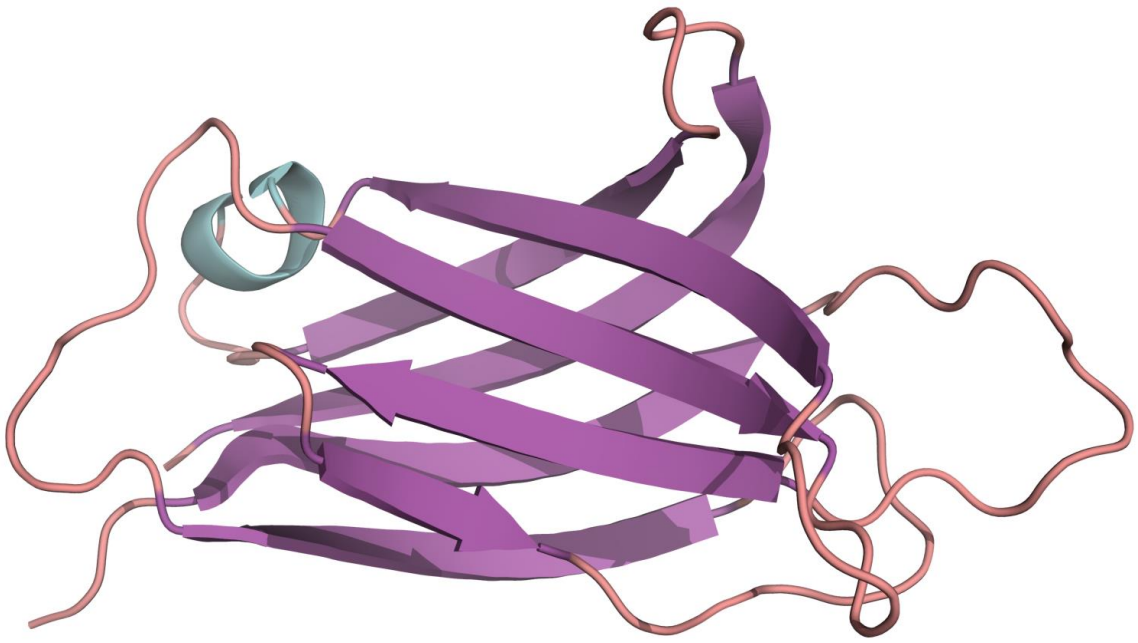

Figure 3 | The solved C2A canonical structure of human dysferlin. C2 is a structural domain implicated in membrane trafficking, fusion phospholipid binding and signaling. It

has a beta-sandwich that is composed of eight beta strands (Sutton, Davletov et al. 1995). At the right end of the beta-sandwich in this representation is the Ca^{2+} binding loop. This 1.76 Å resolution structure was obtained from the RCSB PDB with ID: 4IQH (Fuson, Rice et al. 2014) and visualized with PyMOL (PyMOL).

1.1.6 Dysferlin protein interactors

Dysferlin interacts with several other proteins that have provided insights in the function of dysferlin (Figure 4). A few are described below.

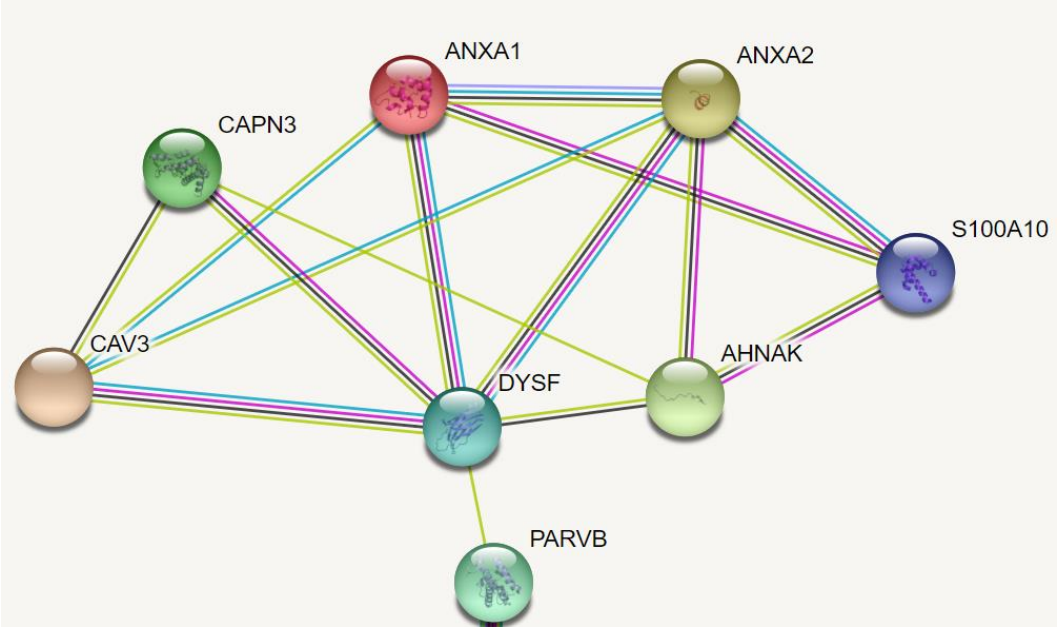

Figure 4 | Dysferlin protein associations. Interactions were obtained from the STRING database (Szklarczyk, Morris et al. 2017). Pink edges represent experimentally determined associations, light blue show associations from curated databases, black show co-expressions, olive green show text-mined associations and purple (*ANXA1* with *ANXA2*) depict protein homology.

Caveolae (CAV) are vesicular invaginations on the plasma membrane with very small size (<100nm in diameter) (Engelman, Zhang et al. 1998). They participate in signal transduction, membrane transport and trafficking, acting as scaffolding proteins for the concentration and organization of specific lipids inside the caveolar membranes

(Galbiati, Razani et al. 2001). Caveolin-3 (CAV3, P56539) is the muscle-specific member of the caveolin protein family. Caveolin-3 is localized to the sarcolemma and is a component of the DGC complex. Defective expression of the CAV3 gene causes LGMD 1C muscular dystrophy (Minetti, Sotgia et al. 1998). Dysferlin was shown to interact with caveolin-3 by coimmunoprecipitation and it was suggested that one of dysferlin functions, in regards to caveolin-3 interaction, is to subserve signalling functions of caveolae (Matsuda, Hayashi et al. 2001). Caveolin-3 deficient patients have a secondary reduction of dysferlin although a converse reduction in caveolin-3 is not always observed in dysferlin-deficiency. When caveolin-3 has defective expression, the localization of dysferlin is abnormal (Matsuda, Hayashi et al. 2001).

Calpain-3 (CAPN3, P20807) is primarily expressed in skeletal muscle and it is the muscle-specific member of the calpain Ca^{2+}-dependant non-lysosomal cystein protease family. It is responsible for LGMD 2A muscular dystrophy (Richard, Broux et al. 1995). Calpain-3 interaction with Dysferlin was shown by coimmunoprecipitation (Huang, Verheesen et al. 2005). Dysferlin deficient patients were also found to have secondary reduction of calpain-3 (Anderson, Harrison et al. 2000). Calpain-3 is thought to be implicated in the muscle membrane repair mechanism because of its interaction with annexins A1 and A2.

Annexins (ANXA1, P04083 and ANXA2, P07355) are ubiquitously expressed Ca^{2+} phospholipid binding proteins that are implicated in signal transduction, membrane trafficking, exocytosis and endocytosis (Raynal and Pollard 1994). Dysferlin Ca^{2+}-dependant interaction with annexins A1 and A2 was shown by coimmunoprecipitation (Lennon, Kho et al. 2003). Also the expression levels of A1 and A2 are higher in dysferlinopathy patients and correlate with the severity of the pathology (Cagliani, Magri et al. 2005), suggesting that A1 and A2 are required in dysferlin-mediated membrane repair in skeletal muscles. McNeil *et al* confirmed the need for A1 in membrane repair (McNeil, Rescher et al. 2006). Annexins are thought to be involved in vesicle to vesicle fusion and movement because they first bind phospholipids, then

initiate vesicle aggregation and finally interact with the actin cytoskeleton (Gerke, Creutz et al. 2005).

AHNAK is a protein family that includes two very large, ~600-700 kDa, proteins that share sequence and structural similarities (Komuro, Masuda et al. 2004): *AHNAK1* (Desmoyokin, Q09666) and *AHNAK2* (Q8IVF2). They are expressed in many cells but have higher expression in skeletal and cardiac muscle cells. AHNAKs are localized in the cytoplasm and nucleus in non-epithelial cells. The C-terminal region of *AHNAK1* and 2 interacts with the C2A dysferlin domain (Huang, Laval et al. 2007). Dysferlin and AHNAK levels are increased and both are relocalized in the cytoplasm during regeneration, suggesting they function together in membrane fusion events.

Affixin (Beta-parvin, *PARVB*, Q9HBI1) is a focal adhesion protein, contains two tandem calponin homology domains and is expressed ubiquitously with higher levels on cardiac and skeletal muscle (Yamaji, Suzuki et al. 2001). It localizes to the muscle plasma membrane and coimmunoprecipitates with dysferlin (Matsuda, Kameyama et al. 2005). The intracellular C-terminal region of dysferlin and the CH1 region of Affixin were found to be binding partners (Matsuda, Kameyama et al. 2005). Affixin expression is reduced at the sarcolemma of dysferlinopathic patients (Matsuda, Kameyama et al. 2005). This interaction could play a role in cytoskeletal reorganization, which is needed for vesicle trafficking on the damaged membrane.

1.1.7 Dysferlin mediated membrane repair

The details of the mechanisms involved in skeletal muscle membrane repair are still unclear, although dysferlin plays a major role in it (Bansal and Campbell 2004; Cooper and Head 2015). Membrane repair requires intracellular vesicles (lysosomes, endosomes, enlargeosomes, etc) to accumulate on the disrupted area and form a patch through Ca^{2+}-dependant vesicular exocytosis (Bi, Alderton et al. 1995; Reddy, Caler et al. 2001; McNeil, Miyake et al. 2003). First the intracellular vesicles are transported to the lesion site via motor proteins such as kinesin, non-muscle myosin IIA and IIB, and MG53 which is a muscle-specific tripartite motif family protein member (TRIM72) (Bi, Morris et

al. 1997; Togo and Steinhardt 2004; Weisleder, Takeshima et al. 2009). Then the vesicles are fused in a Ca2+-dependent manner with the plasma membrane to form a "membrane patch" (Han and Campbell 2007). Vesicle fusion involves synaptotagmins and the SNARE protein family. It is thought that dysferlin acts as a Ca2+ sensor that regulates the SNARE vesicle-membrane fusion, during membrane resealing alongside annexin. Finally, the patch is thought to be removed either by autophagy, endocytosis or phagocytosis by macrophages (Middel, Zhou et al. 2016).

Dysferlin is highly expressed and is associated with the t-tubule network. The network is vulnerable to eccentric stretch, and DYSF-null muscle fibers show t-tubule abnormalities after in vivo lengthening strain injuries similar to those of CAV3-null muscle fibers (Klinge, Harris et al. 2010; Kerr, Ziman et al. 2013), suggesting that dysferlin is important for t-tubule formation and maintenance. Further studies have shown that dysferlin is cleaved by activated calpains after the plasma membrane is injured. The cleavage product, mini-dysferlin$_{C72}$, has only the last C2 domains and the transmembrane domain. These are the most conserved C2 domains and the structure of mini-dysferlin$_{C72}$ resembles those of synaptotagmins (Lek, Lek et al. 2010). Thus it is suggested that the cleaved dysferlin may be recruited to the injury site but not the full length protein. It seems that defective membrane repair could be only one of multiple contributing factors to dysferlin-deficient pathology. It is known that dysferlin deficiency affects trafficking and signaling growth factors (Demonbreun, Fahrenbach et al. 2011) and adhesion molecules (Sharma, Yu et al. 2010) and the late onset of the disease suggests that there must be differences in the need for dysferlin in trafficking and membrane repair between children and adult muscles.

1.1.8 Dysferlinopathies

Dysferlin deficiency in skeletal muscle results in a large variety of muscular dystrophies. Dysferlinopathies are autosomal recessive inherited muscle wasting diseases. The first phenotype was described in 1967 (Miyoshi, Saijo et al. 1967) and in 1986 (Miyoshi, Kawai et al. 1986) by Miyoshi. Subsequently this disorder was called

Miyoshi Myopathy (MM). “Dysferlinopathy” as a term was first mentioned by Kate Bushby when Miyoshi Myopathy and LGMD2B were first found to share the same allele (Bushby 1999; Bushby 1999).

To date, 416 disease-causing mutations have been reported (see http://www.umd.be/DYSF/) (Beroud, Collod-Beroud et al. 2000; Beroud, Hamroun et al. 2005; Blandin, Beroud et al. 2012) in different regions of the *DYSF* gene, but with no mutation hotspots. These include stop codon mutations, frameshifts which lead to premature truncated protein, missense mutations that affect protein stability and deletions (Mahjneh, Bushby et al. 1996; Aoki, Liu et al. 2001; Cagliani, Fortunato et al. 2003; Takahashi, Aoki et al. 2003; Nguyen, Bassez et al. 2005; Therrien, Dodig et al. 2006; Wenzel, Carl et al. 2006; De Luna, Freixas et al. 2007; Krahn, Beroud et al. 2009; Klinge, Aboumousa et al. 2010). Also, several mutations specific to distinct populations (founder mutations) have been reported: Italian (Cagliani, Fortunato et al. 2003), aboriginal Canadian (Weiler, Greenberg et al. 1996), Portuguese (Vernengo, Oliveira et al. 2011) and Palestinian (Mahjneh, Vannelli et al. 1992).

Patients usually have varying symptoms but commonly include slow progressive muscle wasting and weakness accompanied with increased serum levels of creatine kinase (CK) at the early stages of the disease (Urtizberea, Bassez et al. 2008). Typically, the lower limbs are affected prior to upper ones (Mahjneh, Marconi et al. 2001). There are also dysferlin mutations that have no symptoms or with only higher CK levels (HyperCKemia) (Urtizberea, Bassez et al. 2008). Dysferlinopathies in general do not seem to interfere with the respiratory system or cardiac muscles, but studies suggest a mild cardiomyopathy (Wenzel, Geier et al. 2007; Chase, Cox et al. 2009). Dysferlin deficient carriers are usually unaffected, but mild muscle weakness is reported to be present sometimes (Illa, De Luna et al. 2007). The three main clinical phenotypes are: (i) Miyoshi Myopathy (MM) (Miyoshi, Kawai et al. 1986; Bejaoui, Hirabayashi et al. 1995; Liu, Aoki et al. 1998), the distal onset muscular dystrophy; (ii) Limb-Girdle Muscular Dystrophy type 2B (LGMD2B) (Bashir, Strachan et al. 1994; Bashir, Britton et al. 1998), the proximal muscular dystrophy form; and (iii) Distal Myopathy with Anterior Tibial

onset (DMAT) (Illa, Serrano-Munuera et al. 2001), which is very similar to MM except that in the beginning it affects the anterior muscles of the lower limbs. Other clinical phenotypes have also been reported such as proximo-distal weakness (Nguyen, Bassez et al. 2007; Seror, Krahn et al. 2008). A short description of the three main phenotypes follows.

- **Miyoshi Myopathy:** MM or Miyoshi Muscular Dystrophy 1 (MMD1) (OMIM # 254130) (Miyoshi, Kawai et al. 1986; Bejaoui, Hirabayashi et al. 1995; Liu, Aoki et al. 1998) is the most common autosomal recessive myopathy with distal onset and is also the most known type of dysferlinopathy. MM progress is typically slow (decades) and around 15% of the patients will become non-ambulatory. The symptoms usually appear at early adulthood, and include elevated levels of CK and lactate dehydrogenase (LDH), at the early stages of the disease, and muscle weakness that initially begins from the gastrocnemius muscle (calf muscle). Patients usually first report inability to stand on their toes, difficulties getting downstairs and leg pains alongside calf swelling (Diers, Carl et al. 2007). The symptoms, in early adulthood, are quite delayed compared to other early onset muscular dystrophies such as Duchenne Muscular Dystrophy (DMD) (Blake, Weir et al. 2002). Most of the patients have no signs of muscle weakness in their early adulthood. The most notable symptom is the reduced calf size. Over time, muscle wasting extends to the distal upper limb and the pelvic muscles.
- **Limb-Girdle Muscular Dystrophy type 2B**: The clinical phenotype of LGMD2B (OMIM # 253601) (Bashir, Strachan et al. 1994; Bashir, Britton et al. 1998) is very similar with that of MM but, predominantly affects proximal muscles especially quadriceps and hamstrings. The age of onset is on early adulthood and the progression is slow. The shoulder girdle is affected after years of progression have passed.

- **Distal Myopathy with Anterior Tibial onset:** Distal Myopathy with Anterior Tibial onset (DMAT) (OMIM # 606768) (Illa, Serrano-Munuera et al. 2001), is comparable to MM with different affected muscles. It first affects the anterior tibial muscles of the lower limbs and the progresses to the posterior ones. Onset of the disorder is in early adulthood and is rapidly progressive with involvement of the proximal muscles. DMAT is also similar to Nonaka myopathy (Nonaka, Sunohara et al. 1981) since the onset is on the anterior tibial muscles, but with higher CK levels.

1.1.9 Mouse and cell models of dysferlin deficiency

Two mouse strains are typically used to study dysferlinopathy because they each contain a natural occurring dysferlin mutation: SJL/J (JAX 000686) and A/J (JAX 000646) strains (a short description follows). More information about dysferlin deficient mouse models is available at the Jain Foundation website (https://www.jain-foundation.org).

1.1.9a SJL/J and SJL-Dysf

The SJL/J mouse was developed in 1955 at The Jackson Laboratory. It has been reported that the mouse was susceptible to autoimmune disorders and inflammatory muscle diseases (Bernard and Carnegie 1975; Rosenberg, Ringel et al. 1987). It was later shown that the skeletal muscle of SJL/J had increased regenerative capacity compared to BALB/c mouse (Grounds and McGeachie 1989; Mitchell, McGeachie et al. 1992; McGeachie and Grounds 1995). In 1999, Bittner *et al.* uncovered a reduction in dysferlin protein that is consistent with the reduction in dysferlinopathy patients (Bittner, Anderson et al. 1999). A splicing mutation in the 3' splice junction of the *Dysf* gene results in the deletion of exon 45 from dysferlin's mRNA. This is a 171 bp in-frame deletion, removing 57 amino acids (predicted) which belong to the C2E domain of dysferlin (Vafiadaki, Reis et al. 2001) (https://www.jax.org). The *Dysf*^{im} (inflammatory myopathy) allele results to decreased dysferlin protein levels (<15% than the controls).

Mild muscle weakness can be detected histologically at about 3 weeks of age with the main pathology presentation occurring after 6 months by affecting the proximal muscles first (Bittner, Anderson et al. 1999). At 16 months half of the skeletal muscles are replaced by fat tissue (Weller, Magliato et al. 1997). The proximal muscles, quadriceps femoris and triceps brachii, are more severely affected than the distal ones (gastrocnemius, soleus and tibialis anterior) and the progression of the disease is faster than the A/J strain. *Dysf*^{im}/*Dysf*^{im} mutation has been transferred to the C57BL/10 background (C57BL/10.SJL-Dysf, JAX 011128) in order for the C57Bl/10J to be used as an experimental control (https://www.jain-foundation.org). This background exhibits similar characteristics: progressive muscular dystrophy, myofiber degeneration, increased fibrosis and CK levels, which makes it ideal as a model of dysferlinopathy (https://www.jax.org).

1.1.9b A/J and Bla/J

The A/J mouse was first developed in 1921 from a cross between Cold Spring Harbor albino and Bagg albino (Strong 1936) (https://www.jax.org). A unique retrotransposon (6000 bp) is inserted in Dysferlin's fourth intron (5' end) that causes alternative splicing and loss of dysferlin protein (Ho, Post et al. 2004). First symptoms are observed within 4 or 5 months with slow muscle weakness progression with proximal and abdominal being the first affected muscles followed by the distal muscles. In 2010 the *Dysf*^{prmd}/*Dysf*^{prmd} (progressive muscular dystrophy) mutation of the A/J mouse was transferred to the C57BL/6J background, called Bla/J (B6.A-Dysf^{Prmd}Gene/J), making the C57BL/6J mouse a control for experiments (Lostal, Bartoli et al. 2010). The Bla/J mouse exhibits elevated numbers of centronucleated fibers and muscle impairment in most muscle groups within 4 months (disease onset is 2 months). The most affected muscles are psoas, quadriceps femoris, tibialis anterior, and gastrocnemius, in order of severity (Lostal, Bartoli et al. 2010). Reduced membrane repair capability following laser wounding was also shown by Lostal et al (Lostal, Bartoli et al. 2010). New studies reported that the C57BL/6J background results in a more

severe form of dysferlinopathy through increased membrane leakage and inflammation (Demonbreun, Allen et al. 2016). Like the SJL/J strain, the proximal muscles are severely affected while the distal present a milder weakness (abdominal muscles are severely affected too). The progress of the disease is slower than the SJL/J strain.

1.1.9c Cell models

Human immortalized primary myoblasts were isolated from dysferlin patients muscle biopsies and transduced with hTERT and cdk-4 for immortalization while preserving the symptoms of the human dysferlin deficient cells (Philippi, Bigot et al. 2012). Mouse immortalized dysferlin deficient myoblast cells (GREG cells) were derived from the A/J mouse (Humphrey, Mekhedov et al. 2012). These cells also preserve the characteristics of dysferlin deficiency such as, reduced overall dysferlin expression and membrane repair capacity following wounding.

1.1.10 Therapeutic approaches for Dysferlinopathy

There is currently no effective therapeutic option for dysferlinopathy patients. Disease progression is different for dysferlinopathy types and palliative interventions can help: most of the MM patients remain ambulatory throughout their lives, however LGMD 2B patients require a wheelchair within two or three decades after diagnosis.

Since dysferlin deficiency impairs skeletal muscle membrane repair, gene based therapies (gene replacement) that could increase functional dysferlin expression look very promising. Most of these methods use adeno-associated viral vectors (AAVs) to systemically deliver the dysferlin gene while not causing immunological reaction. However, AAVs have a limited size capacity and the dysferlin gene is one of the largest in the human body, thus a truncated and functional version of dysferlin must be generated. One technique is to use a two-vector system to deliver the dysferlin gene in two segments with a large overlapping region and reassemble it inside the cell

(Sondergaard, Griffin et al. 2015). A first phase clinical trial using this method began in 2016 with the participation of 3 low dose and 3 high dose patients of the dual dysferlin AAVs (rAAVrh74.MHCK7.DYSF.DV) on the extensor digitorum brevis muscles. A new study from the same team demonstrated that dysferlin was still expressed in 15 months dysferlin-null mice after the initial injection at 8 weeks old (Potter, Griffin et al. 2017). They also treated DYSF-null mice at a later age, 6 months, which is about the time the phenotype starts to appear, and found improvement on the treated muscles compared to untreated (Potter, Griffin et al. 2017). In a new study a nano-dysferlin, including the important regions, was designed so it could fit inside an AAV and was reported to successfully improve expression levels of dysferlin in Bla/J mice (Llanga, Nagy et al. 2017).

Other therapeutic methods are being investigated such as exon skipping. Exon skipping as a therapy has been developed first for DMD, and is designed to restore the reading frame by removal of an exon adjacent to the deletion site, thereby generating a protein that is truncated but still partially functional. New studies show that certain forms of truncated dysferlin are functional in patient cells (Barthelemy, Blouin et al. 2015). However as there is no mutation hotspot in dysferlin, multiple sites will have to be tested. Furthermore, it is difficult to know which regions are essential for dysferlin to function and if the truncated protein will be functional.

1.2.1 Omics in muscle and neuromuscular pathologies

Omics (also referred as high dimensional biology) approaches aim at the thorough understanding of a complex system by viewing it as a whole. Traditionally, their main goal is to systematically quantify genes, mRNA, proteins, metabolites, etc from a biological sample in an unbiased manner. Omics have revolutionized the field of systems biology (Westerhoff and Palsson 2004). The main difference of systems biology with traditional studies is that the latter are largely hypothesis driven. On the other hand omics experiments are commonly used to generate hypotheses using a holistic

approach, with no prior knowledge or driver, acquiring and analyzing data and defining a hypothesis which can be then tested (Kell and Oliver 2004). Omics approaches are used to understand biological processes but also disease conditions where they can be used for diagnostic or screening purposes. Another way of using omics is for biomarker discovery as they can be used to investigate multiple genes, proteins or molecules at once across multiple conditions. This has led to the currently very popular use of omics technologies for drug discovery assessment and efficacy (Gerhold, Jensen et al. 2002) through the field of pharmacogenomics which could potentially deliver individualized drugs (Evans and Relling 2004).

The Next Generation Sequencing (NGS) explosion has affected all fields of medicine, including neuromuscular disorders, with whole exome, genome or targeted sequencing. NGS can detect genetic single base variations and can therefore be used for diagnostics or gene discovery. NGS diagnostics are strongly applicable to LGMDs as a large number of genes are related with different LGMD subtypes. Also the associated genes have large size and do not have mutation hotspots (Thompson and Straub 2016). Several of the LGMD mutations are rare and usually confined to small populations. Therefore, targeted NGS is used more often as a first diagnosis tool, replacing the single gene methods (Biancalana and Laporte 2015; Thompson and Straub 2016). Instead of immunoanalysis of muscle biopsy and identification of the affected proteins, NGS first methods will use targeted (selected genes after narrowing down the related MDs through clinical screening) or exome sequencing before any further analyses (Lek and MacArthur 2014). If the results show a pathogenic mutation very little follow-up work is required. If they do not give a clear result, then the downstream analyses that were used so far can be used to identify the pathology. Targeted gene sequencing has already been used in many neuromuscular disorders successfully (Ankala, da Silva et al. 2015; Biancalana and Laporte 2015). This can result in associating phenotypes with genes that could not have been tested with the low-throughput methods as they did not seem relevant to the phenotype. Many patients remain with unknown causative mutation(s) even after targeted NGS. In cases like this whole genome sequencing presents a way to

discover genes associated with the phenotypes. In conjunction with NGS, transcriptomic and proteomic approaches are also utilized to evaluate the mutation effect on transcripts and protein levels.

Several large projects aim to understand the genetic causal mutation, underlying mechanisms and ways to develop therapeutic targets in regards to MDs: The European project NeurOmics (http://rd-neuromics.eu) uses omics technologies to develop treatments for 10 neuromuscular diseases, SeqNMD an NIH project that focuses on gene discovery in patients and MYO-SEQ (http://myo-seq.org), a project that collects whole exome sequencing data to patients with unexplained LGMDs. These projects have accumulated more than a thousand LGMD patient sequences so far (Thompson and Straub 2016).

Transcriptomics is the study of the expression levels of total mRNA in a cell or organism. The transcriptome is defined as the genes that are actively expressed at a given moment. One of the first truly high-throughput technologies, the microarray, was developed in the mid to late 90s and was used broadly by researchers for the past 20 years giving new insights on gene functions and interactions. There are many types of microarrays for various biological assays. For example, SNP (single nucleotide polymorphisms) arrays can detect variations in DNA sequences. Gene expression microarrays measure mRNA as gene activity (expression levels). They measure the expression of thousands of genes simultaneously and can analyze the difference of DNA sequence between biological samples.

Although microarrays gave a huge boost in genomics and transcriptomics studies, they have several limitations. DNA microarrays measure changes in fluorescence signal following hybridization to predetermined cDNA probe sequences and by default cannot measure the absolute mRNA abundance (i.e. to compare one probe to another). This became possible with the introduction of RNA-Seq which is the high throughput sequencing of transcript cDNAs. RNA-Seq technology maps the entire transcriptome at an affordable cost. A great wealth of microarray data has been accumulated in public repositories, but in the past few years microarray assays are

gradually being replaced with NGS technology for new studies. NGS can reveal abnormalities such as chromosomal insertions and deletions and measure the absolute expression value of genes simultaneously and more accurately than microarrays, although at a higher computational time cost. NGS technologies have a profound effect on every field of biology and medicine due to the spiraling low cost of the technology, the accuracy and ease of use.

Proteomics is the study of the proteome and its function in a system. The proteome is the set of all expressed proteins in a system at a certain time-point. The goal of proteomics can range from simple cataloging of proteins in a system to more complex study such as quantitative and functional proteomics in different states, and the understanding of this in the context of protein pathways and networks (Larance and Lamond 2015). Proteomics is very promising tool for biomarker discovery since proteins are frequently affected in a disease state, and once identified as biomarkers they can be sensitively assayed using very specific antibody-based detection kits (e.g. ELISA). This is reflected in the many protein disease biomarkers already available (Parker and Borchers 2014). Limitations of proteomics include the inability to accurately detect low abundant proteins and that the approach is very expensive.

Metabolomics is the study of global metabolite profiles in a system (organism or tissue) under a given set of conditions. The metabolome is the complement of all low molecular weight molecules (metabolites) that are present in a specific physiological state. The metabolome is also closer to the phenotype being studied, since metabolites are often the immediate effectors of function. Although the metabolome contains about 5000 metabolites the diversity of the molecules makes it more challenging to assay and to interpret than other omics approaches.

Several aspects of this thesis relate to microarray technology because of their abundance in public data repositories. A short but comprehensive review of this technology, especially for Affymetrix arrays and public repositories, follows.

1.2.2 DNA microarrays

Microarray technology has been developed over the past twenty years and has led to more holistic approaches to cellular activity than are possible by the study of individual biological functions of a few related genes, proteins or cell pathways. The development of this technology has provided new and interesting information, and exponentially increased the available data for the understanding of biological systems. Microarrays, similarly to other high-throughput methods, are important for the full understanding of processes taking place in biological systems, and are complementary to common procedures (Schena 2002). Since their initial application as a new technique for large-scale mapping of DNA and the initial success as transcriptome analysis tools, they are used in many areas, adapting the basic concept and combining it with other techniques.

Each DNA microarray consists of a large number of DNA probe assemblies representing specific genetic loci. The probes are immobilized by covalent bonds on a solid surface (usually glass). In other words, gene detectors are immobilized at specific points on a glass tile, smaller than the human palm, with techniques of modern nanotechnology, and this structure is called a microarray. In addition to gene detection probes, protein probes, tissue fragments, metabolite probes, and the like can also be used. Microarrays allow analysis of gene expression, DNA sequence diversity, protein levels and modification and more, with massive and parallel processing. It is a technology with many applications in areas such as genomics, proteomics, diagnostics, etc. Since early 2000, it has enabled the analysis of the whole transcriptome from a tissue or cell in a single experiment.

A researcher can extract useful information about the biological function of an organism by finding out what genes are induced or suppressed at different cell cycles or developmental stages or in response to environmental stimuli, such as hormone response or high temperature. Groups of genes whose expression increases or decreases under the same conditions, are likely to have associated biological function and perhaps a common adjusting mechanism (Brown and Botstein 1999). They could,

for example, have similar promoter sequences for the same transcription factors. Additionally, a pattern of expression for a specific condition represents a useful reference to characterize similar unfamiliar situations. Gene expression is directly correlated with biological functions and microarrays provide large data sets on diseases, aging, pharmaceutical action, hormonal action, mental illness, metabolism, and many other clinical issues. Microarrays also opened a new road in diagnostic methods, and have become increasingly available to use in laboratories and diagnostic facilities (Schena 1996).

1.2.3 The creation of the modern microarray

In the mid-90's the technology of microarrays was born, as we understand it nowadays (Pease, Solas et al. 1994; Schena, Shalon et al. 1995; DeRisi, Penland et al. 1996; Lockhart, Dong et al. 1996). The forerunner of the microarrays is arguably the colony hybridization method, where DNA was cloned to *E. coli* plasmids plated on agar petri plates covered with nitrocellulose leaves (Grunstein and Hogness 1975). Similar but also different microarray technologies were developed:

- **Spotted arrays:** An array of pins is dipped into wells that contain already synthesized DNA probes and deposits (spots) them to predetermined locations on the array surface (DeRisi, Penland et al. 1996).
- **In-situ synthesized arrays:** Short oligonucleotide sequences (oligos) are synthesized directly onto the array surface using photolithographic (Fodor, Read et al. 1991) (Affymetrix / Nimblegen) or inkjet printing methods (Blanchard, Kaiser et al. 1996; Hughes, Mao et al. 2001) (Agilent Technologies). Since 1995, the Affymetrix company has introduced the GeneChip® array (Lockhart, Dong et al. 1996) with proprietary technology and prohibits the use of such technology by others after patenting. The microarrays from Affymetrix are the most commonly used arrays and were also used in several parts of this thesis because

of their abundance. A section below describes in greater detail about its construction, experimental design and data analysis methods.

- **Self assembled arrays:** A collection of beads that contains a set of diverse oligos is applied to a surface with wells slightly larger than the beads (Michael, Taylor et al. 1998; Ferguson, Steemers et al. 2000; Steemers, Ferguson et al. 2000; Walt 2000). This method was patented by Illumina.

In attempting to analyze the genome of many organisms, the need for a functional study of thousands of genes was born simultaneously. One step in this direction was the recognition of gene expression patterns under physiological and pathological conditions.

1.2.4 Description of experimental process

The experimental process concerns the steps to be followed in conducting a microarray experiment (Figure 5). We will focus on gene profile analysis experiments in this short description.

A microarray experiment is a complex sequence of processes that must be completed successfully to ensure the acceptable quality of the data that will be generated and the conclusions to be drawn. Since the steps are many and complex, the chances for the experiment to fail increase. Thus, we must be cautious about high-throughput microarray data (or any other kind of high-throughput data), although, as the techniques improve, it is easier to get the experiments done correctly and extract safer conclusions.

Microarray experiments are also called modern Northern analyzes. In a Northern analysis, a cell's RNA is isolated and separated by its size, in agarose gel, after electrophoresis on a special surface of nitrocellulose or nylon. The surface is then exposed to a solution of labeled probes that specifically detect RNA molecules and fluoresce.

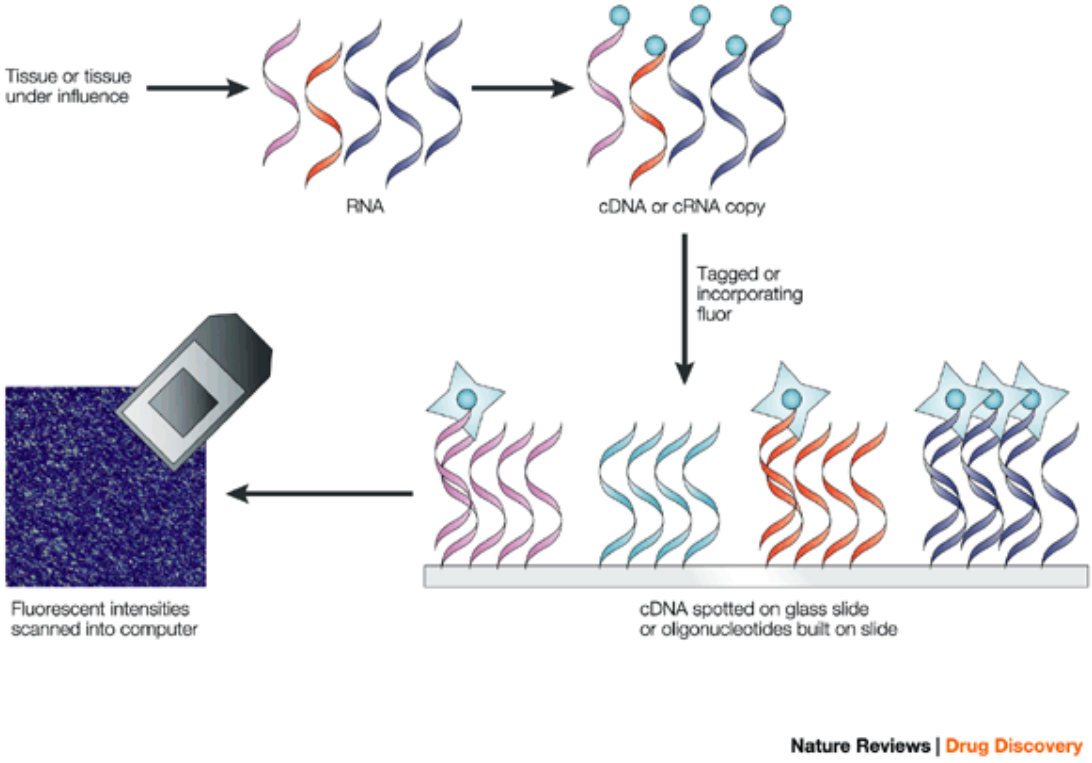

Figure 5 | Experimental procedure in a microarray experiment. Microarray experiments involve isolation of RNA from biological samples of interest, making it fluorescent, hybridize it to the chip, washing off the excess and passing the microarray though a laser light scanner. Image from Butte *et al*. 2002 (Butte 2002).

If we briefly look at a microarray gene profile analysis, initially, we formulate a biological question, which we hope to answer at the end of the experiment. We then proceed to select or construct the microarray, which means what type of microarray will be able to answer the biological question. We then select the microarray in terms of how it is prepared, but also what is the arrangement and type of detectors immobilized on the surface (probes). Probes are the most important selection criterion as they detect the sample's complementary DNA (cDNA) molecules.

At the same time, the biological material is prepared: RNA is isolated from two or more conditions of cells or tissues (e.g. control vs. diseased), if necessary amplified and finally labeled with different pigments for each condition. If using the Affymetrix GeneChip single-channel array (see below), only one condition per tissue can be hybridized in each array. The next step is the hybridization of the fluorescence labeled

target sequences with the microarray probes. Finally, after washing off the excess, we scan the surface of the microarray with a laser light scanner, which returns a digital image derived from the excitation of the labeled molecules found in the target sequences and fluorescing at specific wavelengths.

In a two-channel (two-color) experiment, the target sequences from two different samples are individually stained and hybridized on a single chip. This means that each sample outputs a different colored image at its fluorescent molecule excitation level. If, for example, a control sample is labeled with red fluorescent molecule (Cy5) and a diseased sample with green (Cy3), we are given the possibility of comparing the intensity of the image of each probe after hybridization. Due to the fact that hybridization occurs simultaneously for both samples, there is clear competition between targets for the probes. The labeled target sequences in excess of each sample will bind greater to their corresponding probe. We will see the probe colored red if this gene is overexpressed in the cells of the control sample or, we will see the probe colored green if the corresponding gene for the diseased sample is overexpressed. Finally, the probe will be colored yellow if the expression is similar. In the case that there is no expression we will see the probe colored black (Schena 2002).

When, finally, we obtain the digitally scanned image, we proceed to its analysis, from which quantified data arise. These data are processed with a variety of algorithms to eliminate errors by performing quality controls and to draw conclusions by further downstream analyses: differential expression, clustering, enrichment, correlation, network construction, etc. In the next sections we will focus on the Affymetrix GeneChip array.

1.2.5 Affymetrix GeneChip single-channel oligo arrays

Affymetrix GeneChip arrays consist of monoclonal 25mer oligonucleotide probes, which are synthesized on the solid surface of the microarray by the method of photolithography (Figure 6).

The process begins with the glass plate (wafer), the solid surface of the microarray. The plate is immersed in silane (SiH4) whose molecules are combined with the glass. A linker molecule together with a photosensitive molecule is added to each silane molecule. The linker molecule is the first DNA binding site. The photosensitive molecule acts as a protective molecule (blocker), not allowing new nucleotides to bind to it. A photo mask is placed on the wafer and allows ultraviolet to pass through predetermined points. The exposed spots lose their protection and the wafer is washed with a solution that contains free and single photosensitive modified nucleotides (one at a time). The newly added nucleotides form the substrate where the next ones will bind. This process is repeated until specific 25mer oligonucleotides are formed at every probe location (Lipshutz, Fodor et al. 1999).

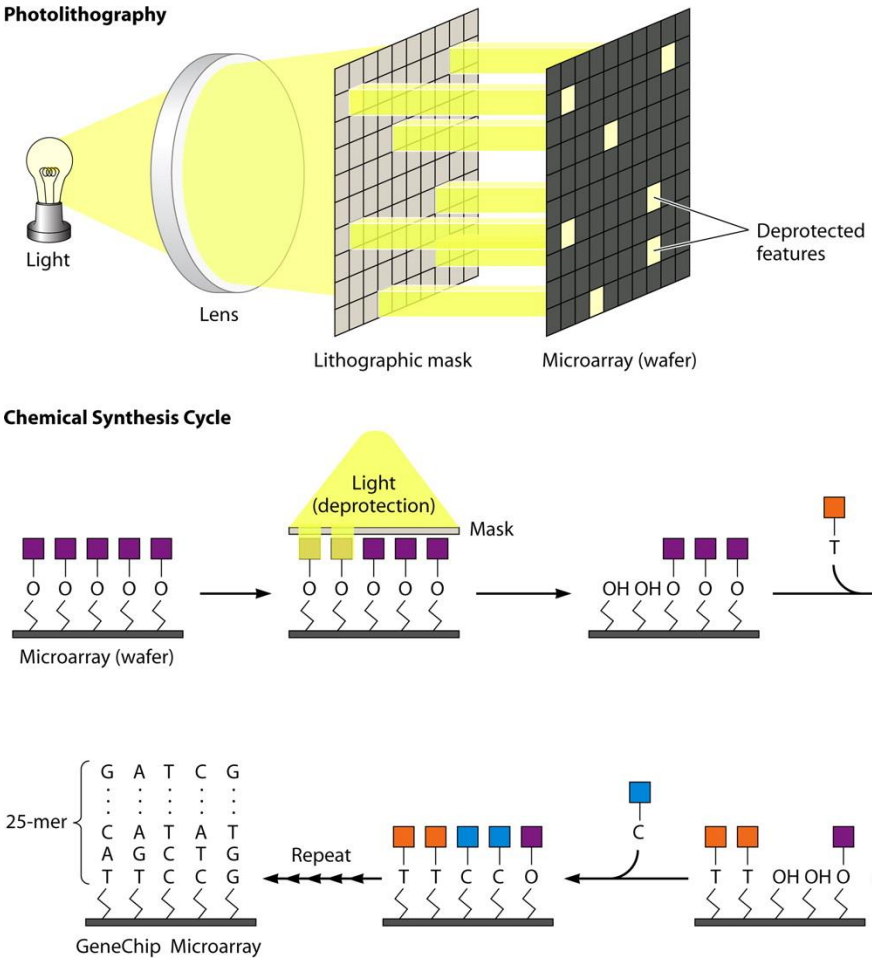

Figure 6 | Affymetrix GeneChip. Top | Photolithography. Ultraviolet radiation passes through the lithographic mask, acting as a filter either to transmit or block the radiation from the chemically protected surface of the microarray. The sequential application of specific lithographic masks determines the order of synthesis of the oligonucleotide probes. **Bottom |** Chemical synthesis cycle. Ultraviolet radiation removes the protective groups from the surface of the microarray, allowing the addition of a single photochemically protected nucleotide. Successive irradiation cycles of de-protection, a change in the filter pattern of lithographic masks, and adding one mononucleotide type at a time, forms microarrays with specific 25mer oligonucleotides-probes. Image from (Dalma-Weiszhausz, Warrington et al. 2006; Miller and Tang 2009).

Each gene or nucleotide sequence is represented by 11 to 20 unique computer-generated probes which are scattered in the microarray to avoid mis-estimation of the quantification of expression due to their location. The probes serve as sensitive, unique and specific sequence sensors. Typically, the probes hybridize to individual regions of the sequence, but sometimes they may overlap a little if deemed necessary. The group of probes specific to a gene or to a similar gene group is known as a probe-set which provides, with high accuracy, the expression value of the target gene. Oligonucleotide probes recognizing parts of the 3' end of the gene are called perfect match probes (PM). The large number of detectors for different regions of the same RNA significantly improves the signal to noise ratio (due to the calculation of a robust mean measure of the intensities of the multiple detection points) and provides precision in the quantification of RNA while preventing cross-hybridization effects and drastically reduces the false positive signals (Lipshutz, Fodor et al. 1999).

Additional quality testing is possible with the use of incomplete match probes (Mismatch or MM). The MM probes have exactly the same nucleotide sequence as the corresponding PM except for the 13th base (middle) which is complementary (Figure 7). The MM probes act as specialized controls that allow direct removal of background and cross-hybridization noise, while distinguishing between true signals and those resulting from non-specific or partial hybridization. Hybridization of labeled RNA sequences in PM produces a higher signal than in MM, resulting in stable patterns that are unlikely to occur randomly. Even at low concentrations of RNA, PM/MM hybridization attributes recognizable patterns which can be quantified (Lipshutz, Fodor et al. 1999). Each MM

detector is located adjacent to that of the corresponding PM to exclude any positional effect.

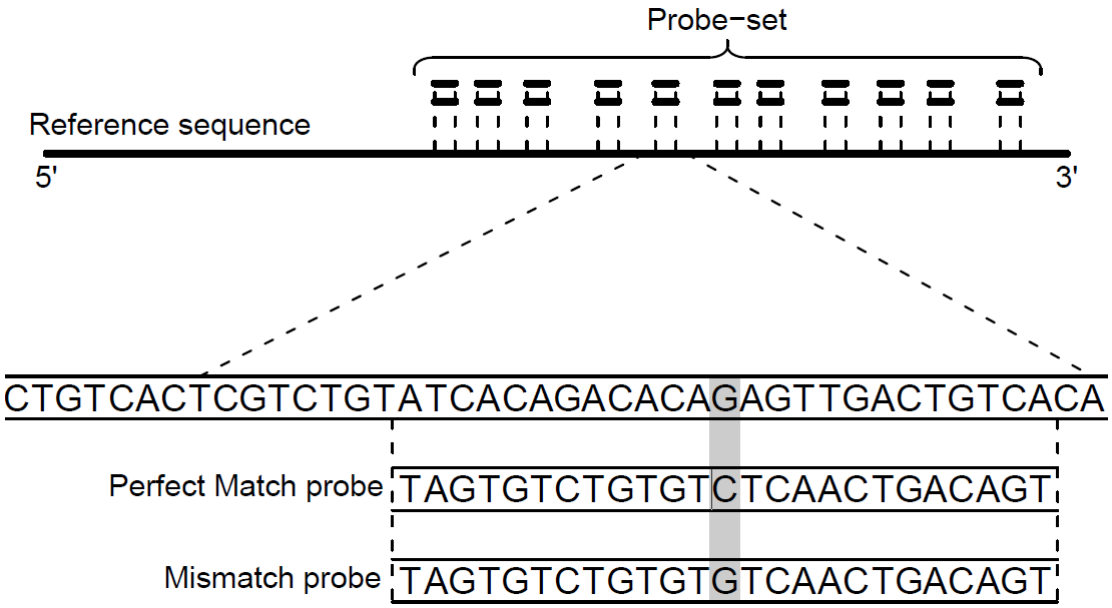

Figure 7 | Affymetrix microarray probe design. Oligonucleotide probes are selected on the basis of uniqueness and design rules. For eukaryotic organisms, the probes are typically selected from the 3' end of the gene or transcript (near the polyA tail), to reduce the problems that may occur from the use of partially degraded RNA. Utilizing the PM difference from MM significantly reduces background and cross-hybridization noise, and increases quantitative precision and reproducibility of the measurements.

1.2.6 Affymetrix GeneChip files

This section describes the most common Affymetrix microarray file formats from raw images to fluorescence light intensities and processed expression values (Figure 8) (Affymetrix 2009):

- **DAT:** Contains the intensity values for each pixel, collected from an Affymetrix scanner.
- **CDF (Chip Description File):** Describes the arrangement of probes on the Affymetrix microarray. A chip usually contains expression, genotyping, specific

labeled and housekeeping probe-sets. All probe-set names within a microarray are unique. Multiple copies of a probe-set may be present, in a chip, if each copy has a unique name.

- **CEL:** Holds the data from each pixel derived from the DAT file and are mostly known as the microarray's raw files. The data include: the light intensity value, the standard deviation of the intensity and the number of pixels used to calculate it. These values are stored for every probe in the array. Two versions of CEL files exist: V3, the text format version and V4 which is the binary version.
- **CHP:** Contains the expression values for each probe-set after background correction, normalization and probe summarization in binary format. The expression values vary based on the preprocessing algorithm that was used. For example, MAS 5.0 exports linear while RMA outputs log2 transformed expression values.
- **TXT:** Usually is the combined probe-set or gene expression matrix in text format, which is used for further downstream analyses (e.g. enrichment analysis).

Initially the DAT file is produced from the Affymetrix scanner and then transformed to the CEL file. Then using the most appropriate CDF in context with the pre-processing algorithm (MAS 5.0, RMA, etc.) the binary CHP file is generated which can then be converted to text and combined with all the samples of the study. Newer versions of pre-processing algorithms export the text file that contains the expression values matrix directly.

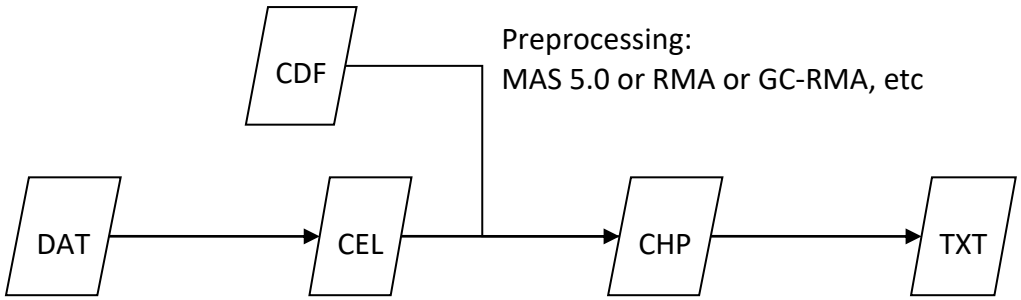

Figure 8 | Common Affymetrix Genechip microarray files.

- DAT: Unprocessed digital image of the hybridized array
- CDF: Chip Description File contains the probes layout within the chip and their assigned transcripts (library is provided by Affymetrix or other sources)
- CEL: Processed DAT file (fluorescence intensity values for each probe and its location)
- CHP: Expression values of each probe-set or gene in binary format after normalization and probe summarization of the corresponding CEL file.
- TXT: Gene expression matrix of every sample in text format.

1.2.8 Microarray data preprocessing

The goal of microarray data analysis is to produce a list of expressed genes for each sample and includes several stages which may vary depending on the type of data analyzed. Prior to any kind of microarray data analysis, several steps are needed to ensure the high quality of the chip. The sample quality and experimental design should be evaluated in order to ensure its integrity. Sources of unwanted variation could be tissue contamination, amount of RNA and its degradation and amplification, reverse transcription and labeling efficiency. Other sources include the DNA quality, PCR yield and cross or unspecific hybridization which may lead to data noise. For some of the variations we can estimate corrections from the data by preprocessing them (Yang Y.H. and P. 2003). Unprocessed raw data are always subject to some form of technical variation and therefore must be preprocessed (often referred to as 'normalization' despite that it includes several other steps) to remove as much as possible undesirable variation to ensure that the results have the highest level of accuracy.

All microarray technologies follow the same general methodology. First, we evaluate the sample quality and experimental design. Then we read the raw data, remove low quality probe-sets or microarrays from further analysis, perform data preprocessing (e.g. RMA (Irizarry, Hobbs et al. 2003)), do more quality controls (Bolstad, Collin et al. 2005; Brettschneider, Collin et al. 2007) on the expression values (preprocessed data) and continue with downstream analyses, such as the calculation of differential expression using appropriate statistics (e.g. t-test, limma (Smyth 2004), etc).

The list of differentially expressed genes then can be supplemented with useful information explaining the function of the various genes, for example, with gene ontology terms (Gene Ontology Consortium 2015) or KEGG pathways (Kanehisa, Furumichi et al. 2017).

Ideally, the data to be analyzed should be preprocessed using various methods, the results of which should be examined to determine which method is best suited (Cope, Irizarry et al. 2004). The most appropriate method should then be used to preprocess the raw data before any further downstream analysis. The next section describes some of the most commonly used preprocessing methods on Affymetrix arrays.

1.2.9 Affymetrix array preprocessing

Due to the design of Affymetrix microarrays, the steps to be taken prior to statistical analysis are slightly more complicated than other cDNA arrays.

1.2.9a Background correction

The first step is the background intensity correction for each probe. The background fluorescence can arise from many sources, such as non-specific binding of the labeled sample to the microarray surface, deposits remaining after the washing step or optical noise from the scanner. Slight fluorescence intensity levels (background noise) will be detected by the scanner, even if only sterile water is labeled and hybridized to the microarray. Preprocessing algorithms use different background correction methods, for example, the RMA algorithm (Irizarry, Hobbs et al. 2003) assumes that PM is a convolution of the true signal (exponential distribution) with the background noise (normal distribution).

1.2.9b Normalization

The next step is normalization. The purpose of this step is to remove the technical variance, while maintaining biological differences between samples. There are always small differences between the hybridization processes for each microarray and these variations tend to lead to large discrepancies between different sample intensities. For example, the amount of RNA in a sample, the time that a sample is hybridized or the volume of a sample, can introduce substantial fluctuations. Even subtle physical differences between or amongst microarray scanners used to scan microarrays, may affect the results.

Simply put, the normalization ensures that the comparison of different microarray sample expression levels is possible. Studies have shown that the normalization methods used play a significant role on the downstream statistic analysis, so it is crucial to choose the appropriate method.

1.2.9c Perfect Match (PM) correction

As mentioned previously, PM probes count both the relative abundance of the corresponding sequence and the amount of nonspecific binding, which occurs when the RNA sequence binds to a probe that should not. The MM probes are designed to count the non-specific binding of the respective detectors PM. MM intensity values should then be subtracted from their respective PM values.

In reality, however, this does not work, because in general about 30% of the MM values are actually higher than their PM counterparts (Naef, Lim et al. 2002; Irizarry, Hobbs et al. 2003). This is because, in addition to the background signal measurement, a significant amount of RNA recognized by the PM, tends to also bind to the MM probes. Many of the most popular preprocessing algorithms solve this problem by simply ignoring the MM probes completely while PM values are corrected for non-specific binding using different approaches (Li and Wong 2001; Naef, Lim et al. 2001; Irizarry, Hobbs et al. 2003).

1.2.9d Probe summarization

We have already seen how the microarray GeneChip operates using 11-20 different PM probes targeting 11-20 nucleotide RNA segments separately. The final step in Affymetrix microarrays data preprocessing is to summarize the intensities from the 11-20 separate probes to one expression value, called probe-set. Several ways are available to achieve this, but the end result is always a unique expression value for each probe-set (Hubbell, Liu et al. 2002; Li 2002; Irizarry, Hobbs et al. 2003; Hochreiter, Clevert et al. 2006; Xing, Kapur et al. 2006).

1.2.10 Preprocessing algorithms for Affymetrix microarrays

Having introduced the general methodology used for preprocessing Affymetrix microarray data, we will describe some of the most popular complex preprocessing algorithms. These algorithms apply all pretreatment steps described above: background correction, expression value normalization for each probe on every microarray, and probe to probe-set summarization. An overall comparison of the preprocessing algorithms can be found at the affycomp website: http://rafalab.rc.fas.harvard.edu/affycomp (Cope, Irizarry et al. 2004; Irizarry, Wu et al. 2006).

1.2.10a MicroArray Suite 5.0 (MAS 5 .0)

MAS 5.0 algorithm was developed by Affymetrix (Affymetrix 2002; Hubbell, Liu et al. 2002; Bolstad, Irizarry et al. 2003; Affymetrix 2004; Gentleman, Carey et al. 2004) and is one of the most widely used single-array method, meaning that it can be computed for each array separately. It consists of 4 steps:

- **Global background correction:** The 2% intensity quantile is subtracted from all probes.
- **Local background correction:** The Ideal Mismatch intensity (IM) is calculated and subtracted from all PM probes. Remember that about 30% of MM probes have higher intensities than their corresponding PM pairs. If the MM intensity is lower

than its PM pair, then IM equals to MM intensity. In the case that MM is equal or greater than the PM, the IM becomes a fraction of the PM intensity.

- **Summarization:** The PM probes are summarized into probe-sets using the one-step Tukey biweight M-estimator.
- **Normalization:** In this step a trimmed mean is calculated, excluding the highest and lowest 2% of the expression values and a target intensity is set (default 500). All expression values are then multiplied by the scaling factor which is the target intensity divided by the trimmed mean value. Therefore, the MAS 5.0 normalizes the data following summarization, not before, as many other algorithms do.

1.2.10a1 Calling absent / present probe-sets

Affymetrix introduced a version of a qualitative expression measurement in the MAS 5.0 algorithm. The reason is that MM probes give a reasonable estimate of the background noise for the majority of the probes in a specific array. So if there is no statistical difference between the PM and MM probe pairs, the gene is considered as non-expressed. As to the accuracy of the MAS 5.0 absent / present call, 85% of the true positive RNA transcripts that are designed to hybridize to control probes, were correctly identified as present (Choe, Boutros et al. 2005). The Wilcoxon's rank test is used for the characterization of a gene as present or absent. This method is used with great success for the quality assessment of the array (McCall, Murakami et al. 2011): Affymetrix recommends a similar percentage of present genes per sample. In the event of uneven percentage of present genes between samples, the sample that has more than 10% difference than the rest, should be considered as a low quality sample.

1.2.10b Probe Logarithmic Intensity Error (PLIER) estimation

PLIER is a multiple array analysis method introduced by Affymetrix, which means it shares information across all samples (Hubbell 2005; Hubbell 2005). It introduces higher signal reproducibility (less variation) without loss of accuracy. It offers higher sensitivity to changes in target abundance near background and dynamically balances

the probes that contain more information from a probe-set to determine the expression value.

1.2.10c Robust Multi-array Average (RMA)

RMA is the most frequently used algorithm to convert probe intensities into gene expression values (Irizarry, Bolstad et al. 2003; Irizarry, Hobbs et al. 2003). As the name suggests, it combines information across the samples, except for the background correction step. This method differs from the Affymetrix methods described above, because it ignores the MM probe values and the normalization step is before the summarization. While recognizing that the MM sensors provide useful information they also introduce noise and at the time of publication of the method, the authors could not find a productive way to use them. A convolution model is used for background correction. It assumes that PM intensity is a sum of background noise and real signal. The corrected PM intensity is the expectation of the real intensity given the total signal. The intensities are then normalized with quantile normalization (Amaratunga and Cabrera 2001). Finally, each probe-set is summarized separately but within all arrays, with the median polish algorithm (Tukey 1977) fitting a two-way ANOVA model. Because the expression values are the estimated array effects, they are in log2 scale.

1.2.10d Gene Chip RMA (GC-RMA)

The GC-RMA is a modification of the RMA algorithm and can only be used in Affymetrix Genechip arrays. In reality, it differs from RMA in the background correction step which makes use of both PM and MM probes to estimate the background better (Wu, Irizarry et al. 2004). It also uses the probe sequence information to detect probe affinity to non-specific binding. This model suggests a probe affinity that is dependent on the position of each base and the base composition of each probe, suggesting that the sequence can affect the intensity of the probe, independent of target concentration (Naef and Magnasco 2003). This leads to improved precision, but at the expense of slightly lower accuracy (detection of relative transcript expression without

concentration bias). It is reported that GC-RMA performs better than the other algorithms on detecting low-intensity, differentially expressed genes (Wu, Irizarry et al. 2004; Schuster, Blanc et al. 2007).

1.2.11 Data standards and data exchange

Microarrays are possibly the earliest biological technology that allowed the collection of vast amounts of digital raw data and processed information. As microarrays gained popularity, a common method that described in detail the microarray chip, the study, its samples, the protocols and the data analysis techniques used, needed to be established so the microarray experiments can be reproduced easily. It also rapidly became apparent that other researchers should have access to raw and processed data that would allow them to (a) perform analyses that the original researchers had not conceived, (b) analyze the data with future state-of-the-art techniques or (c) combine samples from different studies to perform meta-analyses. To overcome these issues, the members of the Microarray Gene Expression Data Society (Brazma, Robinson et al. 2000) (now Functional Genomics Data Society) created the MIAME (Minimum Information About a Microarray Experiment) standards for the description of microarray experiments (Brazma, Hingamp et al. 2001; Ball and Brazma 2006). MIAME is a common language for representing and communicating microarray data. It includes information about the overall experimental design, the design of the microarray (i.e. identification of each probe in each microarray), the origin of each probe and the labeling method, procedures, hybridization parameters and measurement (including the normalization methods). The six most critical MIAME elements are:

- Primary data for each hybridized sample (e.g. CEL or GPR files).
- Preprocessed (normalized) data for all hybridized samples in the experimental study (e.g. the gene expression data matrix used to draw conclusions from the study).

- Basic sample annotation, including experimental factors and their values (e.g. agent and dose in a dose response experiment).
- Experimental design, including relations of samples with data (e.g. which raw data files is associated with which samples, which samples are technical and which biological replicates).
- Sufficient microarray annotation (e.g. gene symbols and names, genomic coordinates, oligonucleotide probe sequences, commercial microarray catalog number).
- Basic laboratory and data processing protocols (e.g. which normalization method is used to obtain the final processed data).

For microarray data exchange, using the unified modeling language (UML), the MIAME metadata were translated to XLM based MAGE-ML and later MAGE-TAB data formats (Spellman, Miller et al. 2002; Rayner, Rocca-Serra et al. 2006). These efforts influenced the creation of data standards in other biological areas as well (Taylor, Paton et al. 2007; Deutsch, Ball et al. 2008; Field, Garrity et al. 2008). As the technology was extensively used, vast amounts of complex transcriptomics data started to accumulate and the need to store and distribute the data gave birth to the two major high throughput genomic databases: GEO (Gene Expression Omnibus) (Edgar, Domrachev et al. 2002; Barrett, Troup et al. 2007) maintained from NCBI and ArrayExpress (Brazma, Parkinson et al. 2003; Brazma, Kapushesky et al. 2006) maintained from EBI. A short description of the GEO database follows. ArrayExpress follows a similar design and definitions.

1.2.12 The GEO repository

The database Gene Expression Omnibus (GEO) (Barrett, Wilhite et al. 2013) of NCBI serves as a public repository for a wide diversity of high-thoughput data. These data include single-channel (Affymetrix GeneChip or Illumina BeadArrays) and two-channel (cDNA) mRNA, genomic DNA, proteins microarrays, and other technologies such

as serial analysis of gene expression (SAGE) (Velculescu, Zhang et al. 1995), mass spectrometry proteomics data and next-generation sequencing (NGS) (Pettersson, Lundeberg et al. 2009). Furthermore, the unprocessed raw data files are almost always deposited alongside the processed high-throughput data.

At the basic organization level of GEO, there are four basic types of entity. The first three (samples, platforms and series) are supplied to GEO by the submitters. The GEO staff assembles and curates the fourth type, datasets, using the data submitted by the users (Figure 9). A short description of the GEO entities follows.

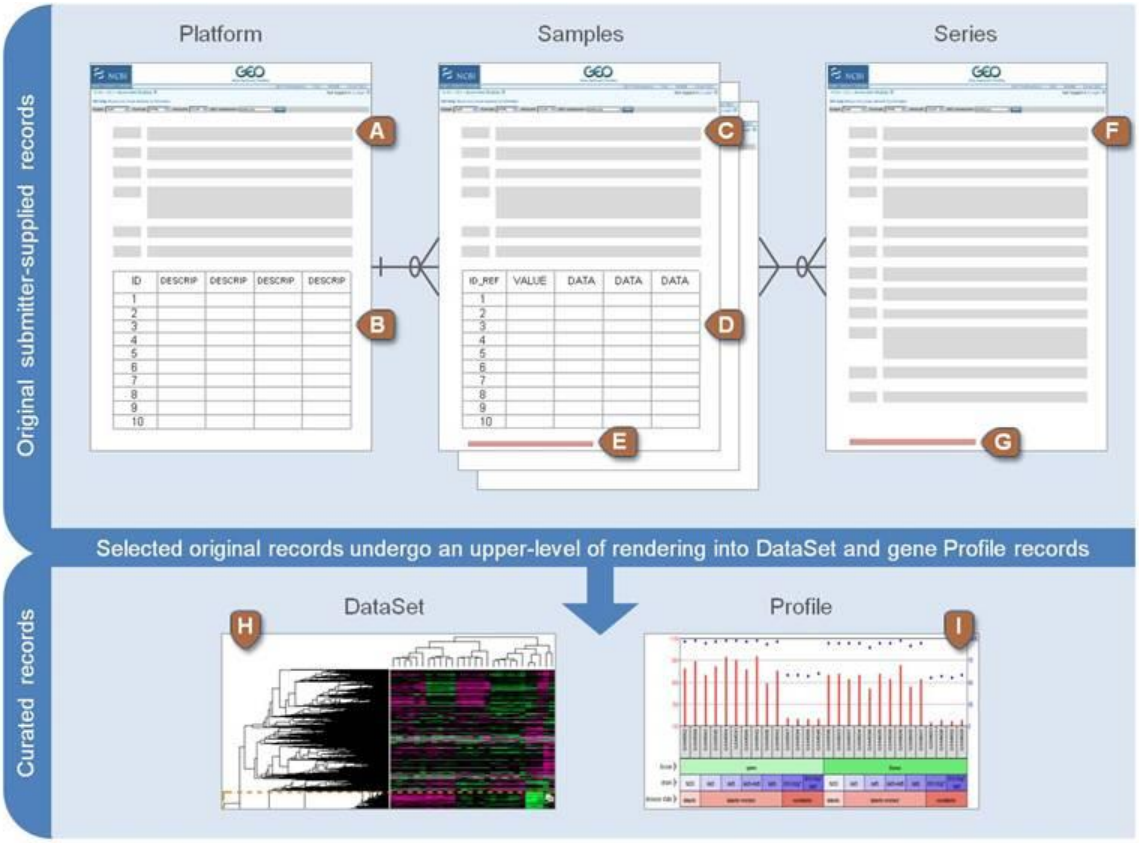

Figure 9 | Representation of database records GEO. A) Description of the microarray. **B)** Table showing the platform model. **C)** Description of the biological sample and protocols incurred. **D)** Sample expression matrix with the processed expression values. **E)** Original raw data file. **F)** Experiment description (Series). **G)** Compressed tar file with the primary values of all samples from that Series. **H)** Datasets have a separate interface with additional computational tools (https://www.ncbi.nlm.nih.gov/sites/GDSbrowser). **I)** Profiles are derived from datasets and consist of the expression measurements for an individual gene across all Samples in a dataset. Profiles can be searched using the GEO

Profiles interface (https://www.ncbi.nlm.nih.gov/geoprofiles). Image from GEO (https://www.ncbi.nlm.nih.gov/geo/info/overview.html).

1.2.12a Platforms (GPL)

A GEO platform record (GPL) describes the features of a microarray chip (e.g. cDNA, oligonucleotide probes, ORFs, antibodies), the list of elements that can be detected and quantified in this experiment (e.g., SAGE signatures, peptides), etc. Each platform record has a unique and stable GEO number and always starts with the letters "GPL" followed by numbers (e.g. The platform GPL 96 describes the Affymetrix Human Genome U133A microarray). The platform may refer to many samples submitted by various users (Figure 10).

1.2.12b Samples (GSM)

A GEO sample record (GSM) describes the origin of each individual sample, the experimental collection, extraction, labeling, hybridization and scanning organization, the computational preprocessing of the primary raw data and the expression value of each probe-set in that sample. Each sample record has a unique and stable GEO number always starting with the letters "GSM" followed by numbers (e.g. the sample GSM845740 is an injured skin biopsy from a patient suffering from psoriasis and was hybridized on the Affymetrix Human Genome U133 Plus 2.0 microarray). Each sample must refer to only a single platform and can be included in one or more series (Figure 10).

1.2.12c Series (GSE)

A GEO series (GSE) record defines a collection of samples that belong to a group (experiment) and explains how the samples are related and arranged. The Series is the focal point of the collection of experimental descriptions. The series documents may also contain tables that describe exported data, summary conclusions or analyses. Each

series record has a unique and fixed GEO number that always begins with the letters "GSE", followed by numbers (e.g. the series GSE34248 includes 28 samples from skin biopsies of patients with and without psoriasis) and may also include samples from different platforms (Figure 10).

1.2.12d Datasets (GDS)

The GEO Datasets (GDS) are curated series or groups of samples. A Dataset record is a collection of biological and statistically comparable samples and is the foundation of the GEO analytical and data display web applications. Samples of each dataset belong exclusively to a platform (Figure 10). The values of each sample that belongs to a given dataset are calculated in an identical manner: factors such as background processing and normalization are common throughout the dataset. Further information, reflecting the experimental design, is provided through dataset subsets. Each dataset record has a unique and fixed GEO number that always begins with the letters "GDS" followed by numbers (e.g. the dataset GDS4100 includes 24 saliva samples from patients with pancreatic cancer and from healthy donors).

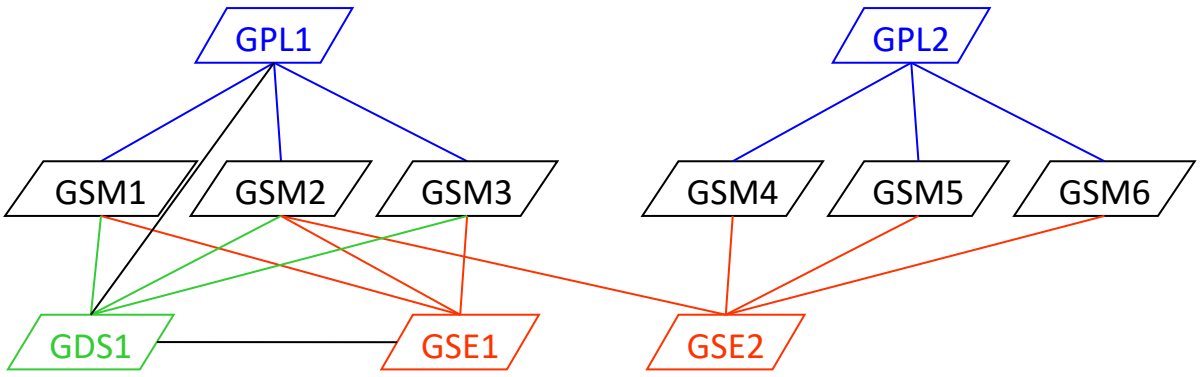

Figure 10 | GEO data structure. Each platform has a unique GPL number. Each sample has a unique GSM number and belongs to only one platform. Each Series has a unique GSE number and it is a set of one or more samples (GSM) that may belong to one or more platforms (GPL). Each dataset has a unique GDS number comprising a set of samples which belong to only one platform and one series.

1.2.13 Data analysis techniques

Described below, are few statistical methods that are used to analyze high dimensional data. These methods are used in general in all types of high-throughput omics data. In biology, they were first used and further developed in the field of transcriptomics, with the rise of microarray technology.

- **Differential expression analysis:** One of the most common downstream types of analysis is the calculation of the differentially expressed genes. First the arrays go through quality controls and after preprocessing an expression matrix with each sample and each gene or probe-set is produced. We can apply the following methods on the gene expression matrix directly.

  One of the simplest methods which is used frequently to rank genes with respect to differential expression, is the fold change. Fold change is the ratio of two means (e.g. diseased/control). The means are calculated for the replicated arrays of each condition. If the values are log transformed, then the ratio is their difference (e.g. $\log_2$(diseased) – $\log_2$(control)). Usually, the genes with fold-change above 2 and below 1/2 are selected. However, the variability of the values is ignored, meaning that genes with high fold-change could also be highly variable and the high fold-change may occur in just one sample.

  Student's t-test has also been used but is also not ideal. Due to the high cost of the experiments, the number of samples is usually small and the variance estimators could appear by chance. Moderated t-tests were developed to improve on the performance of the student's t-test. The empirical Bayes methods (Baldi and Long 2001; Lonnstedt and P. 2002; Kristiansson, Sjogren et al. 2006; Sartor, Tomlinson et al. 2006) modify the variance estimates for more stable results. Alongside the probe-set specific estimators, a global estimator is calculated. Then, weights are computed and used to calculate a weighted mean of the global and probe-set estimators, depending on the variability and

accuracy of the latter ones. Finally, the weighted mean is used as the denominator, instead of the probe-set estimator. The methods of Lonnstedt and Speed (Lonnstedt and P. 2002) were used to develop Limma, one of the most popular empirical bayes t-test package, which is not only used for microarray but also for next generation sequencing data (Smyth 2004; Ritchie, Phipson et al. 2015). Another popular type of moderated t-tests is the Significance Analysis of Microarrays (SAM) method, which adds a constant to the probe-set standard deviation (Tusher, Tibshirani et al. 2001).

- **Dimensionality reduction:** A way to detect non-apparent errors in the experimental data is to use a suitable visualization method. Making a single scatterplot of the data is impossible since each point is highly dimensional. We can explore their relations by dimensionality reduction: instead of having 20,000 dimensions (genes) for each sample, we collapse the information to just 2 or 3 dimensions, while approximately preserving important characteristics such as, the distance between samples. In genomics, the most commonly method used is the linear principal components analysis (PCA) (Pearson 1901). PCA is also used as a dimensionality reduction technique before classification as the principal components maintain the highest variance. Having fewer dimensions with high variance most often increases the accuracy of the classifiers, because the importance of dimensions without variation, that will not help the classifier, is reduced.
- **Clustering:** Unsupervised classification is used to discover whether samples (tissues, conditions, etc.) or genes can be clustered together (Figure 11). It is important to note that sample clustering is different from gene clustering. In the former, tens or hundreds of high-dimensional (described by thousands of genes) samples have to be clustered. In the latter, thousands of genes that are represented by a small number of samples (dimensions) are clustered (D'Haeseleer 2005; de Souto, Costa et al. 2008). Clustering is helpful to determine the relationship of samples and can be used to discover new groups

that were not previously known. Some clustering methods include: hierarchical cluster analysis, k-means clustering (Forgy 1965; Lloyd 1982) and self-organizing maps (Kohonen 1982). These methods require a distance measure between pairs of samples or genes which is usually the Euclidean or Pearson's correlation coefficient distance (Pearson 1920).

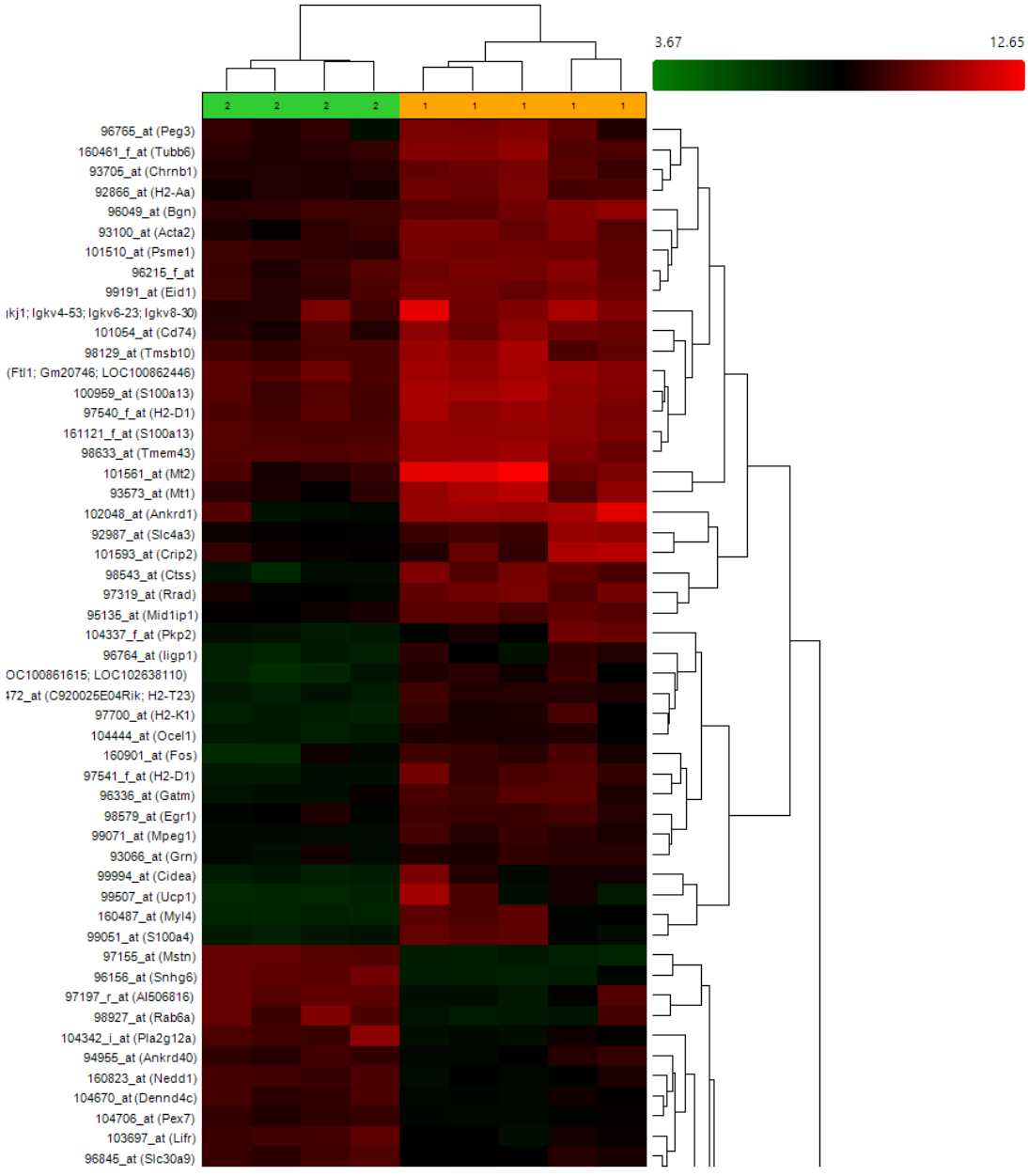

Figure 11 | Part of a typical hierarchical clustering and heatmap between samples and genes. Four dysferlin deficient (bright green) and five control microarray samples (yellow) (GEO series ID: GSE2507) are clustered on the top and are clearly separated. Differentially expressed genes are clustered on the

right. A heatmap is displayed on the center of the image showing the expression levels of the genes (green represents low expression and red high expression).

- **Classification:** Supervised classification is used to train a predictive model to classify future unknown samples into their most likely group/category from the test samples. The input is usually an interesting filtered list of genes derived from other analyses. Classification methods commonly used in genomics are logistic regression, k-nearest neighbor (Altman 1992), random forests (Tin Kam 1998), naive Bayes (Hand and Yu 2001), neural networks and support vector machines (Cortes and Vapnik 1995).
- **Network methods:** Network statistics can be used to represent associative or causative relationships among gene pairs (Emmert-Streib and Dehmer 2008). Gene co-expression networks are often used to identify functional associations of genes "guilt by association", discover hub genes in scale-free topology networks and even correlations between groups of genes that belong to pathways or gene ontology terms (Langfelder and Horvath 2008).

1.2.14 Omics approaches in dysferlinopathy

Since the early 2000s several dysferlin-related omics experiments have already been performed and published in GEO, of which the majority used microarray technology (Table 3). In the course of this thesis, dysferlin-related processed and raw data were collected and analyzed with modern methods and often in a different context from what was intended by the original authors (e.g. combination of the studies or between omics results). We narrowed down the most up-to-date algorithms that are most appropriate for each technology. Each sample was then quality assessed, manually curated and analyzed with modern algorithms, and with identical algorithms if it was from the same technology, so that the obtained information can be more directly compared or combined.

Technology	GEO ID	Samples	Organism / Tissue	Raw data	References
Affymetrix Murine Genome U74Av2	GSE2507	20	Mouse / Mouse left ventricle cardiac muscle, skeletal muscle	Yes	(Wenzel, Zabojszcza et al. 2005)
Affymetrix Murine Genome U74Av2	GSE2629	12	Mouse / quadriceps, tibialis anterior	No	(von der Hagen, Laval et al. 2005)
Affymetrix Human Genome HG-U133A and U133B	GSE3307	28 / 30	Human / skeletal muscle	Yes	(Bakay, Wang et al. 2006)
Affymetrix C. elegans Genome Array	GSE16753	12	C. elegans / adult worm supernatant	Yes	(Krajacic, Hermanowski et al. 2009)
Illumina HumanHT-12 v3.0 Expression BeadChip	GSE26852	11	Human / quadriceps, deltoids	Yes	(Tasca, Pescatori et al. 2012)
Affymetrix GeneChip Human Exon 1.0 ST	GSE44874	11	Human / vastus lateralis myotubes	Yes	(Pakula, Schneider et al. 2013)
Illumina MouseWG-6 v2.0 Expression BeadChip	GSE46420	18	Mouse A/J / tibialis anterior	Yes	(Uaesoontrachoon, Cha et al. 2013)
Affymetrix GeneChip Mouse Exon 1.0 ST	GSE62945	18	Mouse / quadriceps	Yes	(Lee, Lehar et al. 2015)
Spotted oligonucleotide non-commercial LGTCmuOLIs2 (2 channel) (GPL1770). Used Sigma-Genosys oligonucleotide library	GSE2112	4	Mouse SJL/J / quadriceps	No	(Turk, Sterrenburg et al. 2006)
Spotted cDNA non-commercial Human Array 1.0 (GPL2677) (2 channel)	GSE3022	10	Human / skeletal muscle	No	(Campanaro, Romualdi et al. 2002)
Mass spec	NA	2	Mouse Bla/J / skeletal muscle	Yes	unpublished
Mass spec: SILAM LC-MS/MS	NA	45	Mouse Bla/J / quadriceps, tibialis anterior, psoas, gastrocnemius	Yes	unpublished

Table 3 | Dysferlin-related omics data. 10 microarray series from GEO and 2 proteomic studies were analyzed in the course of this thesis.

1.2.15 Objectives

Although omics repositories are accessible to everyone, it is rather challenging for a bench researcher to retrieve raw data, assess their quality and gather the information needed. Also with the amount of raw data produced, methods and tools are required to combine all this information in various ways. Such combination will allow

researchers to identify new functions and interactions or to improve statistical tests and presently undeveloped methods.

For this purpose, we set out to develop bioinformatics tools specific for muscle researchers to have access to omics information without the need of specialized knowledge. The most straightforward way to realize this was to build and maintain the tools as websites. An important requirement for the tools was to bring together information from various resources and databases such as Uniprot and Gene Ontology but oriented towards striated muscle. Finally, we wanted to retrieve and combine all muscle-related available samples from omics repositories and analyze them with state-of-the-art algorithms and most importantly with a robust pipeline that will give us consistent and comparable results.

One of our tools, MyoMiner, retrieves all muscle-related microarray samples that are available in GEO and ArrayExpress in order to calculate the co-expression of expressed gene pairs on muscle tissues and cells with various conditions. From the co-expression matrices, we can then develop networks and use them to functionally associate genes or to interrelate them with other association networks, such as protein-protein interactions (PPI) and pathways. Integrating and curating vast amounts of data can give clearer answers to biological questions. The collection of muscle data accumulated in MyoMiner can complement functional information to muscle-specific genes, create biological networks, identify sets of genes that are regulated on different conditions, and find many more applications.

In the following chapter three publications are provided, describing the MyoMiner and CellWhere tools and a study where dysferlin microarray data were used. A short summary and statement of contribution precedes each publication.

Chapter 2 – Manuscripts

2.1 List of papers and statement of contribution

1. **Apostolos Malatras**, Ioannis Michalopoulos, Gillian Butler-Browne, Simone Spuler, William Duddy. *MyoMiner: A tool to Explore Gene Co-expression in Muscle* (in preparation).

Together with Dr. William Duddy we conceived the project. I assembled and designed a data analysis pipeline, developed analytical tools and constructed the database including the web interface.

2. Aurelia Defour, Sushma Medikayala, Jack H Van der Meulen, Marshall W , ogarth, Nicholas Holdreith, **Apostolos Malatras**, William Duddy, Jessica Boehler, Kanneboyina Nagaraju, Jyoti K Jaiswal (2017). *Annexin A2 links poor myofiber repair with inflammation and adipogenic replacement of the injured muscle*. Human Molecular Genetics. February 21 doi: 10.1093/hmg/ddx065.

Alongside Dr. William Duddy we analyzed the ANXA2 knockout (KO) microarray samples. I collected and analyzed the Dysferlin deficient samples that were compared with the ANXA2 KO samples.

3. Zhu L*, **Malatras A***, Thorley M, Aghoghogbe I, Mer A, Duguez S, Butler-Browne G, Voit T, Duddy W. (2015). *CellWhere: graphical display of interaction networks organized on subcellular localizations*. Nucleic Acids Res. July 1 (* co-first authors) doi: 10.1093/nar/gkv354.

I designed the database, analyzed the Uniprot, GO and Mentha data and implemented the automatic database updates.

The following publications are in Appendix:

4. Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, Jenkins SL, Feldmann AS, Hu KS, McDermott MG, Duan Q, Clark NR, Jones MR, Kou Y, Goff T, Woodland H, Amaral FM, Szeto GL, Fuchs O, Schüssler-Fiorenza Rose SM, Sharma S, Schwartz U, Bausela XB, Szymkiewicz M, Maroulis V, Salykin A, Barra CM, Kruth CD, Bongio NJ, Mathur V, Todoric RD, Rubin UE, **Malatras A**, Fulp CT, Galindo JA, Motiejunaite R, Jüschke C, Dishuck PC, Lahl K, Jafari M, Aibar S, Zaravinos A, Steenhuizen LH, Allison LR, Gamallo P, de Andres Segura F, Dae Devlin T, Pérez- García V, Ma'ayan A (2016). *Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd*. Nature Communications. Sep 26 doi: 10.1038/ncomms12846.

I collected and extracted gene signatures from numerous (mostly muscle) microarray series.

5. Thorley M*, **Malatras A***, Duddy W*, Le Gall L, Mouly V, Butler-Browne G, Duguez S. (2015). *Changes in Communication between Muscle Stem Cells and their Environment with Aging*. Journal of Neuromuscular Diseases. Review (* co-first authors) doi: 10.3233/JND-150097.

I retrieved and analyzed GSE9103 series to answer the question of whether oxidative stress is affected in aged muscles.

2.2 “MyoMiner: A tool to Explore Gene Co-expression in Muscle”

MyoMiner: Explore Gene Co-expression in Normal and Pathological Muscle

Apostolos Malatras[1], Ioannis Michalopoulos[2], Gillian Butler-Browne[1], Simone Spuler[3], William J. Duddy[1,4,*]

[1] Center for Research in Myology 75013, Sorbonne Universités, UPMC University Paris 06, INSERM UMRS975, CNRS FRE3617, GH Pitié Salpêtrière, Paris 13, Paris, France

[2] Centre of Systems Biology, Biomedical Research Foundation, Academy of Athens, Athens 11527, Greece

[3] Muscle Research Unit, Experimental and Clinical Research Center – a joint cooperation of the Charité Medical Faculty and the Max Delbrück Center for Molecular Medicine, Lindenberger Weg 80, 13125 Berlin, Germany

[4] Northern Ireland Centre for Stratified Medicine, Altnagelvin Hospital Campus, Ulster University, Londonderry, Northern Ireland, BT52 1SJ UK

* Correspondence to:

William J Duddy PhD

Email: w.duddy@ulster.ac.uk

Abstract

MyoMiner is a muscle cell- and tissue-specific database that supports co-expression analyses in both normal and pathological muscle tissues. Many gene co-expression databases already exist and are used broadly by researchers but MyoMiner is the first muscle-specific database of its kind. MyoMiner can be accessed at https://myominer-myo.rhcloud.com

High-throughput microarray experiments measure mRNA levels for thousands of genes in a biological sample and most microarray studies are focused on differentially expressed genes. Another way of using microarray data collections is to exploit gene co-expression, which is widely used to study gene regulation and function, protein interactions and signaling pathways.

MyoMiner was created to provide a simple and easy-to-use web interface for muscle scientists to search for transcriptional correlation of any expressed gene pair in muscle cells/tissues and various pathological conditions. We chose the most abundant microarray platforms found on ArrayExpress and GEO repositories, HG-U133 Plus 2.0 for human and MG 430 2.0 for mouse, acquiring 2,376 mouse and 2,228 human samples, and separating them into 142 human, mouse and cell striated muscle categories based on age, sex, anatomic part, and condition. Within each category, users can select a gene of interest, and MyoMiner will return all correlated genes. For each co-expressed gene pair, FDR adjusted p-value and Confidence Intervals are provided as measures of expression correlation strength. A standardized expression-level scatterplot is available for every gene pair's r value. A network tool is also implemented which can be used by the user to create a 2-shell network, based either on the most highly correlated genes, or on a list of genes provided by the user and their correlated or linked genes in the database. Users can also test whether any two correlation coefficients from different conditions are significantly different by using the comparison tool.

These co-expression analyses will help investigators to delineate the tissue-, cell-, and pathology-specific elements of muscle protein interactions, cell signaling and gene regulation. Changes in co-expression between pathologic and healthy tissue may suggest new disease mechanisms and help define novel therapeutic targets. Thus, MyoMiner is a powerful muscle-specific database for the discovery of genes that are associated in related functions based on their co-expression.

Introduction

High-throughput data are an important tool for the study of modern biology. DNA microarrays provide an efficient way to measure the expression of thousands of genes simultaneously (Schena, Shalon et al. 1995; Lockhart, Dong et al. 1996), thus helping the study of fundamental biological processes like gene regulation, signaling pathways and even complex disease traits. The main use of microarrays is differential gene expression analysis where two or more sets of samples are compared (e.g. normal versus treated or diseased) and the up- or down-regulated genes are identified. The integration of large amounts of data over the years on public high-throughput data repositories such as ArrayExpress (Kolesnikov, Hastings et al. 2015) and Gene Expression Omnibus (Barrett, Wilhite et al. 2013) may allow us to identify relations between genes through correlation analysis. However, it is difficult for experimental researchers to extract or combine the information they seek if they have limited bioinformatics expertise.

Correlation data are now widely used to study gene function, protein interactors and biological networks such as signaling pathways (De Smet and Marchal 2010; Marbach, Costello et al. 2012). Furthermore, pathology-specific gene co-expression can be used as a biomarker discovery tool (Sun, Zhang et al. 2014) or for patient prognosis (Futamura, Nishida et al. 2014; Ma, Shen et al. 2014). Several organism-specific co-expression databases already exist such as the Arabidopsis Co-expression Tool (ACT) (Jen, Manfield et al. 2006) and ATTED-II (Aoki, Okamura et al. 2016) for *Arabidopsis thaliana*, and CoXPRESdb (Okamura, Aoki et al. 2015), STARNET (Jupiter, Chen et al. 2009) and Human Gene Correlation Analysis (HGCA) (Michalopoulos, Pavlopoulos et al. 2012) for mammals. They collect gene expression data and a Pearson correlation coefficient is calculated between probes or genes, which can be used as a measure of expression correlation and for network construction from the highly-correlated genes.

However, these databases are not tissue- or cell-specific, because their expression matrices are derived from a mix of tissue types and in some cases from

mixed conditions (e.g. treated and untreated cells). Since gene expression differs between types of tissues and cells (Piro, Ala et al. 2011), it is expected that gene co-expression will also vary. Experimentalists seeking to identify correlation patterns for a chosen gene of interest, usually focus on a specific tissue or cell model and thus the relevance of co-expression values is greatly enhanced by the specificity of the data used (Greene, Krishnan et al. 2015). ImmuCo (Wang, Qi et al. 2015) and Immuno-Navigator (Vandenbon, Dinh et al. 2016) gene co-expression databases are among the first to address immune cell specific correlation, and the latter also correcting the expression matrices for batch effects. Many conditions, such as reagents, equipment, software and personnel, can vary during the course of an experiment and may introduce batch effects, which is a common and strong source of variation on high-throughput data (Leek, Scharpf et al. 2010; Leek 2014). Batch effects are unrelated to biological or scientific variables, are not corrected by normalization (Leek, Scharpf et al. 2010) and must be removed before any further analysis. By combining studies one extra layer of batch effects is introduced: experiments from different laboratories (Irizarry, Warren et al. 2005). If left uncorrected, this technical variation will introduce error into the results of correlation analysis. Another difference of the aforementioned databases is that they only include gene correlation from healthy samples or a mix of healthy and diseased conditions. Studying the changes in correlation between healthy and pathological states could lead to biomarker discovery and to improved understanding of disease mechanisms.

Here, we introduce MyoMiner (https://myominer-myo.rhcloud.com), the first striated muscle cell- and tissue-specific database that provides co-expression analyses in both normal and pathological tissues, addressing both issues of overall correlation and batch effects. MyoMiner includes 2,376 mouse and 2,228 human microarray samples separated in 142 human, mouse and cell categories based on age, sex, anatomic part and condition. We built a simple and easy-to-use web interface to search for transcriptional correlation of any expressed gene pair in muscle cells/tissues and the various pathological conditions. Users can select a category and a gene of interest, and

MyoMiner will return all the expressed correlated genes for that category. Correlation strength is measured by the provided FDR adjusted p-value (q-value) and Confidence Intervals for each correlation.

Materials and Methods

Microarray data collection

Even though ArrayExpress mirrors Gene Expression Omnibus, we searched both repositories for striated muscle (skeletal and cardiac), cells and cell line experiments. In this initial screening we found that the most abundant microarray chips used for muscle related experiments were Affymetrix Human Genome U133 Plus 2.0 GeneChip (GEO platform GPL570 or ArrayExpress ID A-AFFY-44) for human and Affymetrix Mouse Genome 430 2.0 GeneChip (GEO platform GPL1261 or ArrayExpress ID A-AFFY-45) for murine samples. Since a correlation analysis requires very homogenous data, we limited our more refined subsequent searches to these two platforms, which represent about 50 % of all muscle arrays on both repositories.

We searched ArrayExpress using the following string: *(muscle(s) OR myoblast(s) OR myotube(s) OR myofiber(s) OR cardiomyocyte(s) OR myocyte(s) OR heart(s) OR HSMM) AND A-AFFY-44* for human samples and *(muscle(s) OR myoblast(s) OR myotube(s) OR myofiber(s) OR cardiomyocyte(s) OR myocyte(s) OR heart(s) OR C2C12 OR HL1 OR G8 OR SOL8) AND A-AFFY-45* for murine samples. GEO and ArrayExpress assign a different ID (GPL) to each alternative platform. An alternative platform uses the same chip as the original but pre-processed with a different probe-to-gene mapping file called Chip Description File (CDF). It is quite popular for researchers to use a different CDF than the original for better probe-to-probeset and probeset-to-gene targeting accuracy (see "Probes to gene mapping" section). GEO provides a list of alternative platforms in the original platform GPL, but is not well maintained and many are missing. A better way to identify them is to search on ArrayExpress (which is manually curated) for alternative IDs. In the browse page of ArrayExpress* we searched for *U133 Plus 2.0*, MG 430 2.0 and retrieved all the alternative GEO platforms and IDs to A-AFFY-44 (GPL570) for human and to A-AFFY-45 (GPL1261) for mouse (Table S1).

* https://www.ebi.ac.uk/arrayexpress/arrays/browse.html

Next, we parsed their MIAME (Brazma, Hingamp et al. 2001) metadata and confirmed them manually, selecting only those pertinent to muscle research. We excluded all series that did not include the raw CEL files (Affymetrix fluorescence light intensity files), as we pre-processed the CEL files using the robust data analysis pipeline described in detail below, in order to homogenize the data as much as possible.

Particular microarray samples may have been used for several experiments, or analyzed with different normalization algorithms, or even grouped with other samples in big meta-analyses, the results of which have been re-submitted to the repositories. The reused microarrays get a different ID (GSM number in GEO) and it is crucial to identify and remove them from co-expression analysis, as duplicates will erroneously introduce perfect correlation scores. Using the conversion tool (apt-cel-convert.exe) of Affymetrix Power Tools (Affymetrix 2006), we transformed the binary CEL files (version 4) to ASCII text format (version 3) in order to parse them. Their light intensity values (Figure 12) were concatenated into a string and used as input to three hash algorithms: MD5 (Turner and Chen 2011), SHA-1 (Eastlake 2001) and CRC32 (Brayer and Hammond 1975). The hashes act as a unique key for each sample and the duplicate arrays were then easily identified and removed (A simpler version of this algorithm (MD5 on file only) is available in S1).

[CEL]								
Version=3								
[HEADER]								
Cols=712	GridCornerUL=222 233							
Rows=712	GridCornerUR=4484 257							
TotalX=712	GridCornerLR=4460 4527							
TotalY=712	GridCornerLL=198 4503							
OffsetX=0	Axis-invertX=0							
OffsetY=0	AxisInvertY=0							
swapXY=0								
DatHeader=[36..65524] Fusco:CLS=4733.VE=17 08/08/03 11:39:34 HG-U133_Plus_2.1sq								
Algorithm=Percentile								
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.500;OutlierLow:1.004								
[INTENSITY]								
NumberCells=1354896								
CellHeader=X	Y	MEAN	STDV	NPIXELS				
0	0	338.3	39.1	16				
1	0	9772	748.2	16				
2	0	351.5	38.9	16				
3	0	10061.3	817.5	16				
4	0	157.5	20.7	16				
5	0	171.8	26.4	16				
6	0	9317.5	946.2	16				
7	0	188.8	23.2	16				
8	0	9149.5	857.6	16				
9	0	163.8	25.1	16				
10	0	8885.3	886.9	16				
11	0	185.8	25.5	16				

Figure 12 | A typical Affymetrix ASCII text format CEL file. To create a sample-specific hash key we concatenated only the light intensity values (red rectangle) in order to distinguish the unique arrays. It is virtually impossible for different arrays to provide the same intensity values. If the CEL file is in binary format (version 4) we convert it to text format using the Affymetrix Power Tools suit. The processing date of this chip is also visible in row 11 starting with *DatHeader*.

Quality assessment of Affymetrix microarrays

Even though the arrays are published and are thus reported to have passed rigorous quality controls (QC) we performed a global quality control using a battery of BioConductor (Gentleman, Carey et al. 2004; Huber, Carey et al. 2015) packages: 'simpleaffy' (Miller 2017), 'affyQCReport' (Parman, Halling et al. 2017), and 'affyPLM' (Bolstad, Collin et al. 2005; Brettschneider, Collin et al. 2007), using the MAS 5.0 algorithm (Hubbell, Liu et al. 2002) and the Affymetrix default Chip Description File (CDF). We used the Affymetrix chip embedded single array quality metrics for each sample,

such as average background, scale factor, the percentage of genes called present and 3' to 5' RNA hybridization ratios for β-actin and GAPDH. We also used two multi-array quality metrics for each series, Normalized Unscaled Standard Error (NUSE) and Relative Log Expression (RLE). As a general guideline we followed thresholds as recommended by Affymetrix: differences in average background per sample not higher than 20, scale factor within 3-fold change of one sample to another, no higher than 10 % difference of percent present genes and 3' to 5' ratio threshold of GAPDH to 1.25 and β-actin to 3. Also the NUSE boxplots should be centered at 1 with the bad quality ones centered above 1.1. Samples were also deemed as low quality if they had globally higher spread of NUSE distribution than others. Sinse most probes are not changed across the arrays, it is expected that the ratio of probeset expression and the median probeset expression across all samples of a series will be around 0 on a log scale. The RLE boxplots presenting the distribution of these log-ratios should be centered near 0 and have similar spread with low quality samples having a spread higher than 0.2. Arrays that had extreme values or were above our set thresholds on the combined QC's were not used for any further analysis. In total we removed 160 human and 122 mouse samples (Table S2, S3). We identified the poor quality arrays based primarily on the output of percent present, RLE and NUSE, as they are known to perform well (McCall, Murakami et al. 2011), and secondarily on GAPDH and β-actin ratios.

Data normalization

Pre-processing algorithms, usually termed normalization algorithms, are three-step processes: background correction, normalization and probe summarization. The arrays that passed quality controls were pre-processed with the Single Channel Array Normalization (SCAN) algorithm (Piccolo, Sun et al. 2012) with default parameters except for the CDFs, which were downloaded from BrainArray ENSG version 20.0.0 (Dai, Wang et al. 2005). SCAN normalizes each array independently from its series, corrects GC bias and reduces probe and array variation from each individual sample, while increasing signal-to-noise ratio. Single array normalization is preferred when combining

microarray samples from different series or laboratories, because other pre-processing algorithms such as RMA (Irizarry, Hobbs et al. 2003) or GC-RMA (Zhijin Wu 2004) use information across samples for both normalization and summarization steps, and could introduce correlation artifacts (Lim, Wang et al. 2007; Usadel, Obayashi et al. 2009).

Probes-to-genes mapping

The microarray Affymetrix GeneChips we used for MyoMiner are the most abundantly used chips for human and mouse microarray experiments. However, their selection of probes relied on early genome and transcriptome annotation which is significantly different from our current knowledge. The genes on the microarray chips are usually represented by multiple probesets and, conversely, in many cases a single probeset could target multiple genes. Multiple probesets targeting the same gene could exhibit wildly different expression levels making downstream analysis challenging. *Dai et al.* (Dai, Wang et al. 2005), had observed this limitation and created the BrainArray portal where they reorganize probes with up-to-date genomic, cDNA and single nucleotide polymorphism (SNP) information in order to create a more accurate and precise CDF. This has become very popular amongst researchers (Sandberg and Larsson 2007). BrainArray's CDF is updated annually with most microarray algorithms and tools now supporting its CDF by default. The SCAN normalization algorithm has in-built parameters to download and use BrainArray CDFs. For MyoMiner we used Ensembl genome (Aken, Ayling et al. 2016) (ENSG) version 20.0.0. We set the SCAN CDF specified parameter *probeSummaryPackage* to *InstallBrainArrayPackage("human_sample_name.CEL", "20.0.0", "hs", "ensg")* and *InstallBrainArrayPackage("mouse_sample_name.CEL", "20.0.0", "mm", "ensg")* for human and mouse organisms respectively.

Filtering and annotation of expressed genes

In order to distinguish between expressed and unexpressed genes, but also to remove genes with expression levels close to or lower than the background noise, we

used the Universal exPression Code (UPC) algorithm (Piccolo, Withers et al. 2013) separately for each category. We did that because different tissues, cells or pathological conditions have distinct genetic profiles. UPC is a 2-step algorithm that corrects for background noise using linear statistical models and estimates the percentage of gene expression by calculating the active and inactive gene population. An assumption is made that genes with identical molecular characteristics should share the same background expression levels. To identify expressed genes for each category, we calculated UPC's percentage expression 3rd quartile for each gene and categorized it as being expressed if its value was higher than 50 %.

To map Ensembl gene IDs to gene symbols, Entrez IDs (Maglott, Ostell et al. 2011) and Uniprot accession numbers (The UniProt Consortium 2017), we used Ensembl BioMart (Kinsella, Kahari et al. 2011). We extracted the required information from GRCh38.p5 assembly for human and GRCm38.p4 assembly for mouse.

Gender prediction

On half of the MIAME metadata entries for both organisms, the gender information was missing (Florez-Vargas, Brass et al. 2016). To predict the missing gender entries we used hgfocus.db (Carlson 2016) and mouse4302.db (Carlson 2016) from Bioconductor to map genes to chromosomes and then we calculated the median expression of Y chromosome genes. Males should have higher expression values than females, which was visible on the Y chromosome gene expression histogram with two clearly separated gender peaks.

Batch effects evaluation

For batch effect reduction we used the ComBat algorithm (Johnson, Li et al. 2007) from the "SVA" Bioconductor package (Leek, Johnson et al. 2012). ComBat is a robust empirical Bayes method that adjusts for known batch covariates. By default, we used each series as a different batch for every category (gender, age, etc). However, it is

also known that processing time can be a strong batch surrogate (Leek, Scharpf et al. 2010). From the ASCII converted CEL files we retrieved the scan dates (Figure 12 row 11) and used them as batch surrogates for each series, assuming that microarray experiments performed on the same day belonged to the same experimental batch, thus subdividing the aforementioned default series batches to date and series batches. Using principal component analysis (PCA) 3D plots, by the "rgl" R package (Adler, D. et al. 2016), for each category, we identified if the samples correlate with batch surrogates and proceeded with batch correction if necessary. We did not use the category differences as input for the ComBat algorithm (*modcombat=model.matrix(~1,numbatches)*), because a) all samples were from the same category and b) samples that are assigned to a batch are usually unevenly distributed which can induce incorrect differences (Nygaard, Rodland et al. 2016). In some cases, when a batch was represented by a single sample, after assessing the PCA 3D plot we assigned the sample to the closest batch cluster if possible, otherwise we used the *mean.only = TRUE* parameter in ComBat that corrects only the mean of the batch effect not adjusting for scale.

Gene expression correlation

Spearman's rank correlation (Spearman 1904) is a non-parametric rank statistic that measures the strength of a monotonic, linear or non-linear, relationship between two sets of data. Monotonic is a function that increases when its independent variable increases, having a positive correlation. If the independent variable decreases while the function increases, the correlation will be negative. Spearman's correlation is a simply the application of Pearson's correlation (Pearson 1920) on rank converted data. A faster method to calculate Spearman's *r* is to rank the values of x_i and y_i, and calculate their difference d_i. The rank correlation can then be computed as follows:

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \quad \text{(Equation 1)}$$

where "n" is the number of samples and $d_i = rank(x_i) - rank(y_i)$. Spearman's correlation assumes values between -1 and +1, where -1 describes a perfect monotonically decreasing relation and +1 a perfect monotonically increasing relation. If the data are monotonically independent, Spearman's *r* is equal to 0. However, this does not necessarily mean that the data are independent in other ways.

Since Spearman's correlation can be asymptotically approximated by a *t*-distribution with *n-2* degrees of freedom under the null hypothesis of no correlation, we used Student's *t*-test to examine whether a correlation was significantly different from the null hypothesis:

$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{(Equation 2)}$$

To adjust for multiple testing we used the Benjamini – Hochberg (BH) method (Benjamini and Hochberg 1995) to control the false discovery rate (FDR). Correlation *r* and adjusted *p* values were computed with the "psych" R package (Revelle 2017).

Because the correlation coefficient is not distributed normally and its variance is dependent on both sample size and the correlation coefficient from the entire population *ρ*, we cannot compute confidence intervals directly for the *r* values (Lu and Shen). First we have to convert *r* values into additive quantities with *r* to *Z* Fisher transformation (Fisher 1915) which is the inverse hyperbolic tangent function (arctanh):

$$Z_r = \frac{1}{2}\ln\left[\frac{(1+r)}{(1-r)}\right] = \text{arctanh}\,(r) \quad \text{(Equation 3)}$$

its standard error is given by

$$SE_Z = \frac{1}{\sqrt{n-3}} \quad \text{(Equation 4)}$$

where *ln* is the natural logarithm. Second we compute the confidence intervals as follows:

$$CI = Z_r \pm \frac{Z_{table}}{\sqrt{n-3}} \Rightarrow \begin{cases} CI_{upper} = Z_r + \frac{Z_{table}}{\sqrt{n-3}} \\ CI_{lower} = Z_r - \frac{Z_{table}}{\sqrt{n-3}} \end{cases}$$ (Equation 5)

at 95% confidence level $Z_{table} = 1.96$. The final step is to convert *Z* scores back to *r* values using the hyperbolic tangent function (tanh):

$$r = \frac{1-e^{2Z}}{1+e^{2Z}} = \tanh(Z)$$ (Equation 6)

where *e* is the natural base. So in any sample correlation coefficient *r*, there is a 95% probability that the true population correlation coefficient value *ρ* will be in the range of CI_{lower} and CI_{upper}.

For comparing whether any two correlation coefficients r_1 and r_2, for different categories (various samples and sample sizes), are significantly different, we make the null hypothesis (H_0) that the correlation coefficients are not statistically different. Then we transform the *r* values to *Z* scores

$$Z_r = \frac{1}{2}\ln\left[\frac{(1+r)}{(1-r)}\right] = \operatorname{arctanh}(r)$$

(Equation 3), calculate the difference between them and calculate an absolute *Z* score by dividing the difference with the pooled standard error:

$$Z_c = \left|\frac{Z_1 - Z_2}{SE_{zp}}\right|, \text{ where } SE_{zp} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$ (Equation 7)

If $Z_c < Z_{table}$ where $Z_{table} = 1.96$ or more commonly $Z_c > 0.05$ since Z_c is reported as p-value on MyoMiner, we cannot reject H_0. The difference between r_1 and r_2 is not significant at 95% confidence level.

Database construction and website implementation

We developed an easy to use HTML5 web portal that allows querying and visualizing for the requested gene correlations. The interface was developed using the Bootstrap responsive framework. Scatterplots and correlation networks are visualized with the NVD3.js and D3.js javascript libraries respectively. All Spearman's rank and its p values pairwise matrices, and metadata are stored on a relational MySQL database management system which runs on the Apache web server. Dynamic content is processed by the PHP programming language: data retrieval, r to Z transformations and CI calculations. The front-end is powered by Openshift and the back-end by Okeanos cloud services.

Results

Data statistics

MyoMiner was constructed in several steps using various tools and processes (Figure 13). Initially we intended to populate MyoMiner with the most extensively used microarray chips worldwide: Affymetrix Human Genome U133 Plus 2.0 (GPL570) and Affymetrix Mouse Genome 430 2.0 (GPL1261). After screening for muscle related experiments, these chips remained the most popular, accounting for about half of the muscle microarray experiments, in both human and mouse organisms, which had raw CEL files deposited on GEO or ArrayExpress public repositories. We kept only the experiments with raw CEL files, as we wanted to check for their quality and pre-process all collected samples with the same algorithm and parameters.

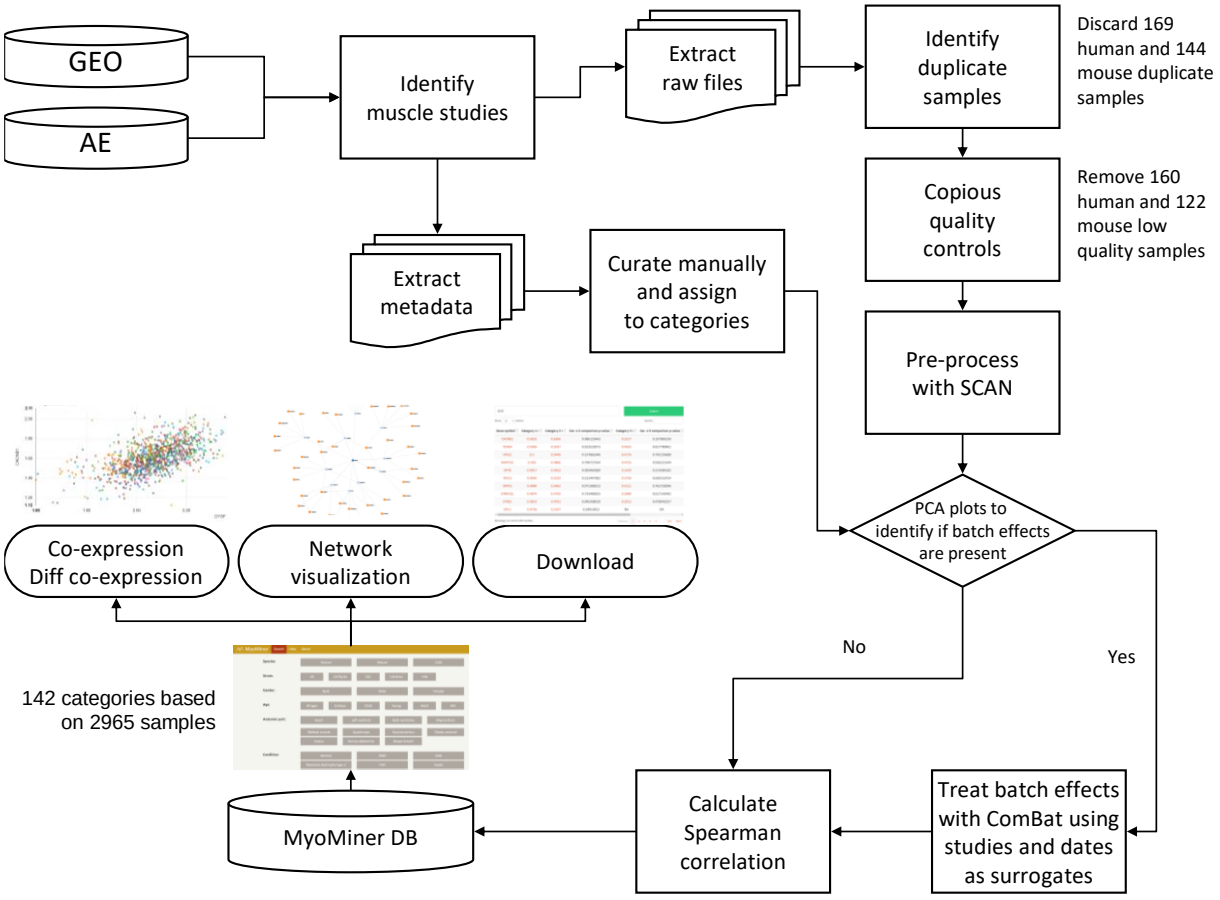

Figure 13 | Workflow of data pre-processing method used for MyoMiner. We identified studies that are pertinent to muscle research from GEO and ArrayExpress. Only the studies that provided the raw CEL files proceeded to quality controls. Samples that passed QC were pre-processed with the SCAN algorithm. We thoroughly curated the metadata files and separated them into categories. We used PCA to detected and remove batch effects using the ComBat algorithm. Users have access to muscle tissue and cells gene-pair co-expression, differential co-expression of every category and co-expression networks. All data are available on the MyoMiner web portal.

Using the advanced search option on the ArrayExpress repository, we filtered and programmatically retrieved 81 human (2541 samples) and 198 mouse (2642 samples) muscle series. We manually parsed each series MIAMI compliant SDRF (sample and data relationship format) metadata file while crosschecking them, if applicable, with the corresponding SOFT (simple omnibus format in text) file from GEO. If there were

missing data or differences between ArrayExpress and GEO we tracked the publication that described the series to correct the missing information. If we still could not extract the missing data, we came in contact with the corresponding authors in case they could provide us with the correct data. Being in close co-operation with ArrayExpress and GEO personnel we corrected several series metafiles, although the most common mismatches were copying errors.

We identified and removed 169 human and 144 mouse samples as duplicates. Finally, 160 human and 122 mouse samples did not pass quality controls and were discarded, leaving us with 74 human series (2228 samples) and 189 mouse series (2376 samples). The samples were then assigned to different categories excluding those that had less than 12 samples. In total 1810 human samples were assigned to 69 categories and 1155 mouse samples were assigned to 73 categories (Table S4).

Categories were created based on gender, age, muscle tissue, condition and strain. A total of 7 skeletal and cardiac muscles tissues are included on MyoMiner together with the combination of those. Human age was classified in years as follows: 0 to 14 as child, 15 to 24 as young, 25 to 59 as adult and 60+ as elderly. For mouse the classification is in weeks: E (embryonic days) as embryo, 0 to 11 weeks as young, 12 to 24 as adult and 25+ as old. We also included 4 separate strains for mouse: C57BL/6J, CD1, C3H/HeJ and FVB but also the combinations of them and more strains (Table 4). Cells are derived from mouse microarrays: skeletal muscle precursor cells and cardiomyocytes, but also from the immortalized C2C12 mouse cell lines in different stages of differentiation: myoblasts, myotubes 1-2, 3-4 and 5+ days after differentiation. MyoMiner covers 53 distinct conditions including normal and pathological ones. In detail, several exercise categories: aerobic, resistance, endurance, trained or sedentary, different types of diets: high fat or calorie restricted diet, type 2 diabetes (DM2): Pre-DM2, DM2 relatives, etc, muscle regeneration: cardiotoxin and glycerol injections, several cardiomyopathies: Idiopathic, Dilated, Ischemic and Arrhythmogenic, muscular dystrophies: Duschenne muscular dystrophy, Mdx, Myotonic dystrophy type 2 and many more (Table S4, MyoMiner web portal).

Organism	Human					Mouse				
Gender	Both		Male		Female	Both		Male	Female	
Age	All ages	Child	Young	Adult	Old	All ages	Embryo	Young	Adult	Old
Anatomic part	Combined heart	Left ventricle	Both ventricles	Myocardium		Combined heart		Left ventricle	Both ventricles	
	Combined skeletal muscle	Quadriceps	Rectus abdominis	Biceps brachii		Combined skeletal muscle	Quad-riceps	Gastro-cnemius	Tibialis anterior	Soleus
Strain	NA					Combined	C57BL/6J	CD1	C3H/HeJ	FVB

Table 4 | Gender, age, tissue and strain classification for each organism. 7 distinct muscle tissues, 4 different age stages (years for human and weeks for mouse) and 4 separated mouse strains with their combinations.

To measure the accuracy of the gender prediction method we first tried it on the samples with known gender. For human only 1135 out of 2228 samples had their gender reported. The method classified 98% of the samples correctly to their respected gender. 23 samples (~2%) did not match and we investigated further into their original publications. We then identified and corrected 5 samples out of 23 which were predicted as opposite sex incorrectly and increased the initial accuracy to 98.4%. For mouse the gender was known in 1390 out of 2376 samples. Again testing this method on the known gender samples resulted in about 98% accuracy, with 56 samples being predicted as opposite sex from the ones reported. We identified and corrected 16 mis-predicted cases and increased the prediction accuracy to 98.3% (Table S5). All gender mismatches that we corrected occurred from copying errors.

Query results and features

MyoMiner was designed as a simple, easy-to-use and understand website that users could search and immediately retrieve the transcriptional co-expression of any expressed gene pair in muscle tissue and cells. All categories are presented as buttons on the main page (Figure 14 A). When selecting a category, the options that are not relevant to it are deactivated, in order to help the user with the remaining options. MyoMiner supports queries using gene symbols, Ensembl IDs (e.g. ENSG00000135636), Entrez gene IDs (e.g. 8291) and Uniprot accession numbers (e.g. O75923). The table output retrieves the correlation values for all expressed gene pairs in the selected category (Figure 14 B) sorted by r-value. The first column comprises the paired gene symbols which can also be clicked to search for their list of correlated genes. The second column is a description of the paired gene, also serving as a link to the associated gene on GeneCards (Safran, Dalah et al. 2010). The third column shows the Spearman's correlation coefficient but also if clicked the scatterplot of this pair. The fourth and fifth columns report two statistic summaries for the user to judge the significance of the correlation: the BH FDR adjusted p-value and, the CI at 95% confidence level that include information about the estimated effect size and the uncertainty associated with this estimate. CI translates to 95% probability that the population correlation coefficient true value ρ is between CI_{lower} and CI_{upper}. A search bar is provided on the top right corner of the table output for easy gene pair finding and the columns can be sorted by clicking on their headers (e.g. sort by positive or negative correlation). The table can be downloaded, in various formats, using the buttons at the bottom left corner.

Scatterplots are important as supplementary information to help interpret the correlation coefficient. In MyoMiner, interactive expression scatterplots for any gene pair can be accessed by clicking on the r value. A modal window will appear showing the normalized expression values obtained by SCAN for the selected gene pair (Figure 14 C). The series that were used for the selected category are displayed at the top of the scatterplot. By clicking or double-clicking the series ID, one can either remove the

selected series or retain that series only, respectively. Removing series on the scatterplot window will not affect the r value as it is pre-computed for all series shown on the scatterplot.

Correlation networks can be accessed by selecting the network tab and pressing the submit button without the need to re-select the category (Figure 14 D). A signed un-weighted 2-shell network will be constructed. It works either with the number of co-expressed genes in each shell (default: 15 and 5 genes for 1st and 2nd shell respectively) or by setting a correlation threshold through the advanced options. A combination of these two methods is also possible.

Another feature is the gene list network, available through the advanced options, where the user can input a list of genes to create the correlation network. In this case, default 1st and 2nd shell values are set to 0 in order to firstly identify if the genes on the list are related. These values can be changed by the users need. The search form “Locate genes in the network” will hide for a short time all the genes in the network except for the searched gene, making it easy to pinpoint the location of genes inside the network. The link threshold bar can be used to remove edges below a certain correlation value, creating sub networks in the process. The blue colored node is used to point the queried gene, the light blue depicts the 1st shell connected nodes and orange the 2nd shell nodes. Users can pan and zoom by click-dragging on an empty space of the interactive network area and using the mouse wheel, respectively. The nodes are interactive and can be moved to any space of the network area. Users can also double-click a node to highlight its immediate connected nodes.

Since correlation networks can grow quite large, including thousands of nodes and many more edges, it could take several minutes to retrieve the values for large networks from the database. For this reason, we decided that network construction will be a client side task, using the D3 javascript library. For large networks, we recommend using the Chrome browser as it could take some time to render big networks, especially

on low end machines. We also recommend having the graphics card enabled for the browser in order to avoid lag on the rendered network.

Differential co-expression analysis is emerging as a method to complement traditional differential expression analysis (Kostka and Spang 2004; de la Fuente 2010). It can detect biologically important differentially co-expressed gene pairs that would otherwise not be detected via co-expression or differential expression (Hudson, Reverter et al. 2009). Differentially co-expressed genes between different conditions are likely to be regulators, thus explaining differences between phenotypes (Li 2002). MyoMiner provides differential co-expression analysis for any gene pair from any category combination. In the "Compare gene co-expression" form, users can set the categories for comparison (Figure 14 E). The first category is compared to the rest after the gene in question is selected. The output includes the gene symbol and its description, the r_1 value from the first category, the r_2 value from the second category and the p-value of the comparison. If the p-value is higher than 0.05 the difference of r_1 and r_2 is not significant at 95% confidence level. MyoMiner supports multiple simultaneously comparisons.

Figure 14 | How to browse MyoMiner. **(A)** *Select a category of interest*. All categories are visible at the beginning, so that the user can find with ease what is available on MyoMiner. By clicking on a category only the options that are related with this category will remain visible. This way the user is guided to the available MyoMiner category. **(B)** *Table output*. Search by gene symbol, Ensembl or Entrez gene ID. All transcriptional co-expressions of any expressed gene-pair displayed when hitting submit. The first column

is the paired gene symbol, the second is the annotation of the paired gene, the third is the Sprearman's correlation of that pair, the fourth and fifth are the BH FDR adjusted p-value and the confidence intervals. The table can be downloaded in CSV format or copied directly to the clipboard **(C)** *Gene pair scatterplot*. The expression values of every sample of the selected category for that gene pair are plotted by clicking on the r value. Each series is shown at the top and can be toggled to display the expression values for any series independently. **(D)** *Correlation network*. The network is constructed based on gene correlation. Users can change the number of relations or set a correlation threshold from the advanced options. **(E)** *Differential co-expression analysis*. Select two or more categories and compare the first to the rest. A gene may be a regulator if its co-expression is significantly altered (p-value) between pathological conditions. MyoMiner can be accessed at https://myominer-myo.rhcloud.com

Improved combined data quality after the correction of batch effects

By combining data from different data sets and laboratories from around the world we introduce unwanted technical variation which needs to be corrected. Another source of strong non-biological variation, we also observed through PCA plots, was the different chip processing days (Leek, Scharpf et al. 2010). To improve the quality of the co-expression values obtained from tens to hundreds of samples, we check each category for the presence of batch effects by different series and/or processing dates. To acquire the scan dates from the microarray CEL files, we parsed them in text format. We then used PCA to visualize the samples from each category, colored by series or processing dates, on a 3D plane (Figure 15 B), in order to identify underlying batch effects. When we observed non-biological variation we corrected it using the ComBat algorithm (Johnson, Li et al. 2007), as described in the Methods section.

Below, we present two examples where batch effect treatment drastically altered the correlation coefficient between the gene pairs (Figure 15). Dysferlin is a type II transmembrane protein that is enriched in skeletal and cardiac muscle and involved in membrane repair (Han and Campbell 2007). Mutations or loss of *DYSF* gene lead to muscular dystrophies called dysferlinopathies. Synaptopodin 2-like (*SYNPO2L*) protein is an important paralog of Synaptopodin-2 (*SYNPO2*) that is involved in active binding and bundling and associated with Duchene muscular dystrophy and myofibrilar myopathy 2.

We selected the adult human resistance exercise category to illustrate how batch correction removes bias introduced when combining data. Before correction, no strong correlation is observed between *DYSF* and *SYNPO2L*: *r = -0.05* (Table 5, also shown with Pearson's correlation coefficients). Clustering and PCA plots show that the samples are grouped by series which may indicate bias (Figure 15 A, B left). The *DYSF* and *SYNPO2L* gene expression scatterplot reveal the extent of the batch effect: even though individual series (different colors) have clear positive correlation the overall correlation is cancelled out when combined (Figure 15 C). In detail the selected category is comprised of three series. Individual series Spearman correlation is GSE47881 *r = 0.6*, GSE48278 *r = 0.3* and GSE28422 *r = 0.67*. We can also average the correlation values using *r*-to-*Z* Fisher's transformation

$$Z_r = \frac{1}{2}\ln\left[\frac{(1+r)}{(1-r)}\right] = \operatorname{arctanh}(r) \quad \text{(Equation 3)}$$

to convert the non additive *r* values to *Z* scores, then average the *Z* scores and finally convert the mean *Z* back to *r* value

$$r = \frac{1-e^{2Z}}{1+e^{2Z}} = \tanh(Z) \quad \text{(Equation 6)}.$$

DYSF-SYNPO2L average Spearman r for the category is 0.54. After we treated the samples with ComBat which reduced the aforementioned bias (Figure 15 A, B, C right) the correlation value increased to 0.62 which could indicate a possible functional association between *DYSF* and *SYNPO2L* (Assadi, Schindler et al. 2008).

In another example between *DYSF* and Synaptopodin (*SYNPO*), which may be modulating actin-based shape and mobility of dendritic spines, we find that batch effects correction reduces the bias inflated correlation *r = 0.62*. Individual series correlation is as follows: GSE47881 *r = 0.31*, GSE48278 *r = -0.4* and GSE28422 *r = 0.64.* The scatterplot also reveals that the series have mixed correlations (Figure 15 D left) and the overall *r* is biased when we combined the series. The average correlation of the three series is 0.21. After removing the bias (Figure 15 D right) the correlation is reduced from 0.62 to 0.36. Gene pairs that had reduced correlation after batch treatment were

more common which indicates that batch correction could reduce the false positive correlations.

	DYSF - SYNPO2L		*DYSF - SYNPO*	
	Spearman r	Pearson r	Spearman r	Pearson r
Untreated	-0.05	0.02	0.62	0.53
Batch treated	0.62	0.65	0.36	0.42
GSE47881	0.6	0.67	0.31	0.39
GSE48278	0.3	0.31	-0.4	-0.08
GSE28422	0.67	0.79	0.64	0.71
Average	0.54	0.62	0.21	0.38

Table 5 | Examples of gene pairs correlation changes after batch treatment. We illustrate two correlation examples i) between *DYSF* and *SYNPO2L,* where the correlation increases significantly and ii) between *DYSF* and *SYNPO,* where the correlation decreases. Both Spearman and Pearson's correlations are available to indicate that batch effects are prevalent in both parametric and non-parametric statistics. We see big changes on their combined correlation coefficients, which is due to the correction of the variation between studies having been done in different labs by different people. In the case of *DYSF - SYNPO2L,* originally there seems to be no correlation on the combined samples, whilst calculating the correlation on the individual series we see a strong positive correlation. This bias is removed after treating for batches with ComBat, resulting in a positive correlation. The example of the *DYSF – SYNPO* pair shows an initial strong positive correlation, while the individual series have mixed positive and negative correlations. Once the bias is removed we see a reduced correlation.

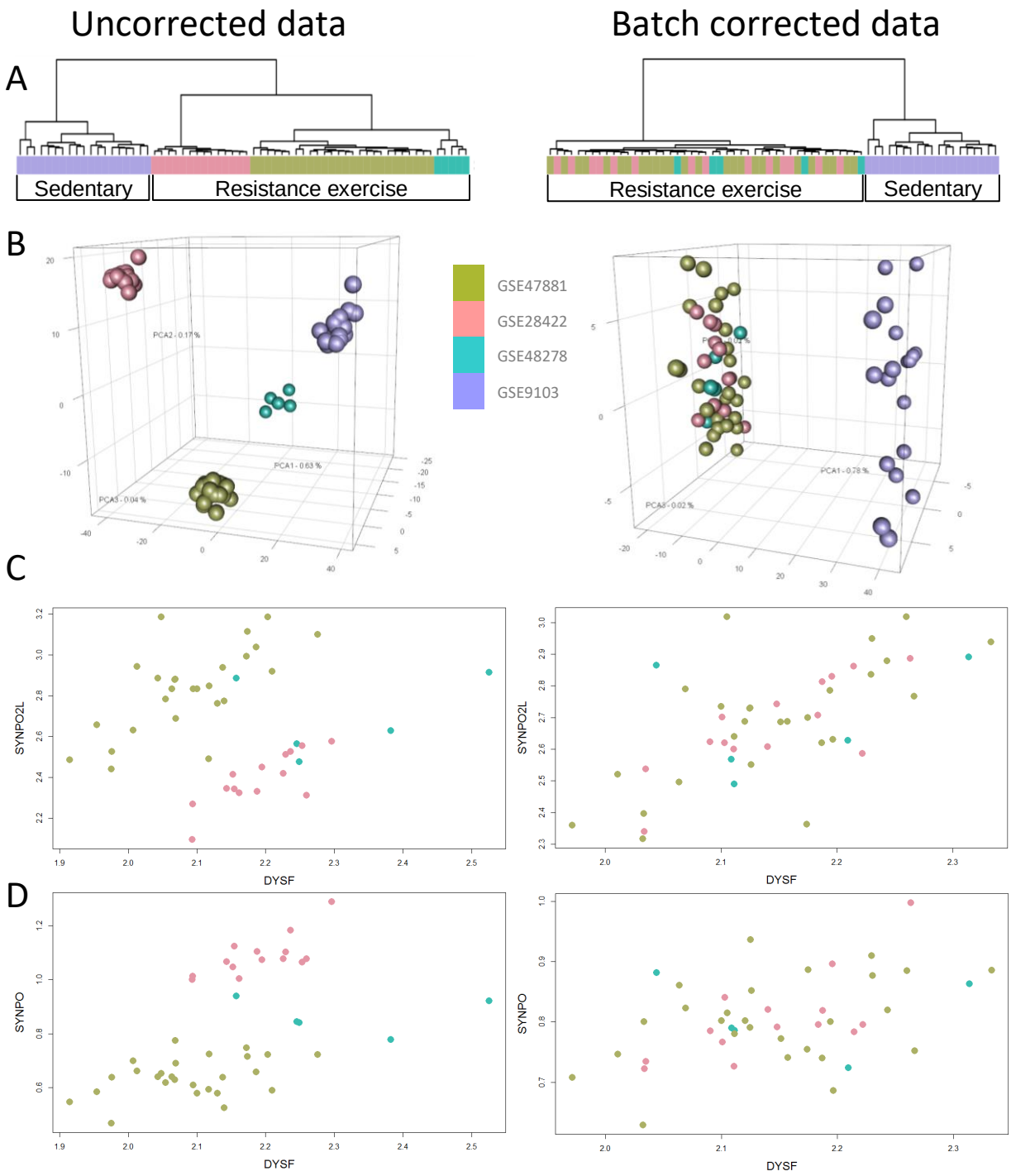

Figure 15 | Example of batch effects treatment. The adult human quadriceps resistance exercise category is constructed from three series: GSE47881 (olive green), GSE28422 (pink) and GSE48278 (turquoise) that include 45 samples in total. GSE9103 (magenta) series, from sedentary individuals, is used as a visual control. On the left, one can see the untreated samples and on the right the batch-treated samples, using each series as a surrogate. **(A)** Hierarchical clustering of both resistance exercise and sedentary

samples shows a clear separation. Note that resistance exercise samples are clustered by their corresponding series even after pre-processing (normalization). After treating the samples with ComBat, the resistance exercise samples are now mixed, reducing the batch effect. **(B)** Principal component analysis plots of the same samples. In the untreated plot, samples are clustered very well by their series (olive green, pink and turquoise). However, the resistance exercise series are as far from each other as the sedentary (visual control in this case) series. After the batch correction (right) all resistance exercise samples are clustered together and are clearly separated from the sedentary samples cluster. **(C)** The expression values of *DYSF* and *SYNPO2L* are grouped by series resulting in a correlation value *r=-0.05*. After batch correction the samples are mixed with *r=0.62*. **(D)** Inversely, in the example of *DYSF* and *SYNPO* where the *r* value is artificially high, before the treatment (r=0.62), the correction reduces it to r=0.36.

Discussion

Gene expression profiling is the most common type of omics data. In this project we retrieved and analyzed striated muscle pertinent microarray samples and combined them effectively for the construction of a muscle-tissue-specific co-expression database. MyoMiner provides a simple, effective and easy way to identify co-expressed gene pairs under a vast number of experimental conditions. This was not available in any other existing co-expression database. Thus, MyoMiner represents a powerful tool for muscle researchers, helping them to delineate gene function and key regulators.

For MyoMiner we chose to use the Spearman correlation coefficient, despite the fact that Pearson correlation seems to be more popular in other correlation databases. We did not use the Pearson correlation because it is sensitive to outliers and because of the assumptions that need to be met, in order to calculate adjusted p-values: every gene would have to be normally distributed, while gene pairs have to be bivariately normally distributed. On the other hand, Spearman correlation is robust to outliers and does not require assumptions of linearity. To determine the strength of the correlation we have provided the adjusted p-value and the confidence intervals, although in cases of many samples in a category, we do not recommend using the arbitrary chosen 0.05 q-value cut-off, but a more stringent value, e.g. 0.005.

It is noteworthy that the most correlated genes for a driver gene may vary significantly between co-expression databases. This can be attributed to different microarray data, although most of these databases use GPL570 and GPL1261 platforms as we did. Moreover, different pre-processing methods, batch effect correction methods or the lack thereof, tissue- and cell-specific expression, variable cell states, different correlation coefficients, etc, add to the differences found in co-expression databases. An investigation of the inconsistencies between co-expression databases could identify common gene characteristics or the key factors that contribute to those differences.

We intend to incorporate GO annotations, protein-protein interaction data from IntACT (Orchard, Ammari et al. 2014) and Mentha (Calderone, Castagnoli et al. 2013), and KEGG pathways (Kanehisa, Furumichi et al. 2017) to further enrich MyoMiner's content. Furthermore, we plan to create three condition-dependant categories; one for cardiac muscles, one for skeletal muscles and one for muscle cell samples. The idea is to include as many different conditions as possible, using a balanced number of representative samples from each condition. This analysis can be used as an initial screening and will help us identify underlying gene pair relationships independent of phenotypes, ages or muscle-tissue type. Since we have a baseline of muscle data and co-expressions, we aim to include more microarray platforms and even RNA-Seq data, so as to include as many neuromuscular disorders as possible.

Supplementary information for MyoMiner

Table S1 | Alternative IDs to the originals A-AFFY-44 for human UG-U133 Plus 2.0 and A-AFFY-45 for mouse MG 430 2 arrays. These experiments get an alternative ID even if they use the same chip because they mapp the probes to probesets and then to transcripts or genes with a different Chip Description File (CDF) than the original. For the muscle microarray collection we pinpointed and downloaded three more series for human and two for mouse.

Affymetrix GeneChip Human Genome U133 Plus 2.0 alternative ArrayExpress IDs			Affymetrix GeneChip Mouse Genome 430 2.0 alternative ArrayExpress IDs	
A-GEOD-4454	A-GEOD-10184	A-GEOD-16268	A-GEOD-5008	A-GEOD-14657
A-GEOD-4866	A-GEOD-10274	A-GEOD-16273	A-GEOD-5759	A-GEOD-14661
A-GEOD-5760	A-GEOD-10335	A-GEOD-16311	A-GEOD-5766	A-GEOD-14757
A-GEOD-6671	A-GEOD-10371	A-GEOD-16356	A-GEOD-6456	A-GEOD-14996
A-GEOD-6732	A-GEOD-10881	A-GEOD-16372	A-GEOD-6526	A-GEOD-15041
A-GEOD-6791	A-GEOD-10925	A-GEOD-17175	A-GEOD-6886	A-GEOD-15592
A-GEOD-6823	A-GEOD-11084	A-GEOD-17180	A-GEOD-7368	A-GEOD-15722
A-GEOD-6879	A-GEOD-11364	A-GEOD-17392	A-GEOD-7546	A-GEOD-15967
A-GEOD-7566	A-GEOD-11433	A-GEOD-17394	A-GEOD-7635	A-GEOD-16225
A-GEOD-7567	A-GEOD-11670	A-GEOD-17810	A-GEOD-8059	A-GEOD-16368
A-GEOD-7869	A-GEOD-13232	A-GEOD-17811	A-GEOD-8462	A-GEOD-16582
A-GEOD-8019	A-GEOD-13668	A-GEOD-17929	A-GEOD-8492	A-GEOD-17109
A-GEOD-8542	A-GEOD-13695	A-GEOD-17996	A-GEOD-9523	A-GEOD-17114
A-GEOD-8715	A-GEOD-13916	A-GEOD-18121	A-GEOD-9746	A-GEOD-18078
A-GEOD-9099	A-GEOD-14837	A-GEOD-18478	A-GEOD-10288	A-GEOD-18122
A-GEOD-9101	A-GEOD-14877	A-GEOD-18850	A-GEOD-10369	A-GEOD-18223
A-GEOD-9102	A-GEOD-15308	A-GEOD-19109	A-GEOD-10773	A-GEOD-18376
A-GEOD-9419	A-GEOD-15394	A-GEOD-19171	A-GEOD-11044	A-GEOD-18416
A-GEOD-9454	A-GEOD-15445	A-GEOD-19883	A-GEOD-13502	A-GEOD-18615
A-GEOD-9486	A-GEOD-15676	A-GEOD-19918	A-GEOD-13621	A-GEOD-18854
A-GEOD-9987	A-GEOD-16006	A-GEOD-20182	A-GEOD-13763	A-GEOD-20766
A-GEOD-10175	A-GEOD-16100	A-MEXP-2335		

Table S2 | Samples and series removed from the human microarray data collection because they failed quality controls. In total 160 samples were considered of low quality and were not used for any further analyses. All samples are removed from the gray shaded series.

Series	Sample	Reason
E-GEOD-1145	GSM18435_PA-D_93_2.CEL	NuSE above 1.1 and RLE wider than +-0.2
E-GEOD-12486	GSM313633.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342678.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342677.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2

E-GEOD-13070	GSM342673.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342808.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342814.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342821.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342836.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342850.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342857.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342879.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342884.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342888.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342900.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13070	GSM342931.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13205	GSM333440.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-15090	GSM377469.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-18715	GSM464627_C12.CEL	RLE wider than +-0.2
E-GEOD-19420	GSM482956.CEL	RLE wider than +-0.2
E-GEOD-22435	GSM557526.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-24199	8 samples	Whole series due to Low percent present and other fluctuations
E-GEOD-24235	GSM596038.CEL	Low percent present compared to other samples from the same series
E-GEOD-24235	GSM595901.CEL	Low percent present compared to other samples from the same series
E-GEOD-25462	GSM624971.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-25462	GSM624970.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-25462	GSM624938.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-28392	70 samples	Whole series due to actin3/actin5 ration being 3 times higher than the recommended limits and RLE widen than +-0.2
E-GEOD-28422	GSM702359.CEL	RLE wider than +-0.2
E-GEOD-28422	GSM702374.CEL	RLE wider than +-0.2
E-GEOD-28422	GSM702438.CEL	RLE wider than +-0.2
E-GEOD-28422	GSM702442.CEL	RLE wider than +-0.2
E-GEOD-34111	GSM842037.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2. Also actin3/5 and gapdh3/5 higher than recommended values
E-GEOD-34111	GSM842028.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2. Also actin3/5 and gapdh3/5 higher than recommended values
E-GEOD-34111	GSM842024.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2. Also actin3/5 and gapdh3/5 higher than recommended values
E-GEOD-34111	GSM842022.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2. Also actin3/5 and gapdh3/5 higher than recommended values
E-GEOD-34111	GSM842018.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2. Also actin3/5 and gapdh3/5 higher than recommended values
E-GEOD-34111	GSM842017.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2. Also actin3/5 and gapdh3/5 higher than recommended values
E-GEOD-34111	GSM842014.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2. Also actin3/5 and gapdh3/5 higher than recommended values

E-GEOD-3526	GSM80797.CEL	Low percent present compared to other samples from the same series. RLE wider than +-0.2
E-GEOD-3526	GSM80796.CEL	Low percent present compared to other samples from the same series. RLE wider than +-0.2
E-GEOD-38780	GSM949395_AA12_15_11_D8.CEL	RLE wider than +-0.2
E-GEOD-39454	GSM969502_MA45_GEIM385.CEL	RLE wider than +-0.2
E-GEOD-39454	GSM969496_MA45_GEIM375.CEL	RLE wider than +-0.2
E-GEOD-39454	GSM969489_MA45_GEIM354.CEL	RLE wider than +-0.2
E-GEOD-40231	GSM988933_STAGE_9_SKLM.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-40231	GSM988889_STAGE_59_SKLM.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-40231	GSM988877_STAGE_56_SKLM.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-40231	GSM988762_STAGE_31_SKLM.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-40231	GSM988759_STAGE_30_SKLM.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-45426	GSM1104107_S26.CEL	RLE wider than +-0.2
E-GEOD-45426	GSM1104095_S14.CEL	RLE wider than +-0.2
E-GEOD-47874	GSM1161401_75_51545Pre.CEL	RLE wider than +-0.2
E-GEOD-47881	GSM1161775_D4_BH073F.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.4
E-GEOD-47881	GSM1161834_D79_PC035F.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.4
E-GEOD-47969	GSM1163791_DUKE38_334.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.4
E-GEOD-48278	GSM1174154_MJH_STRRIDE_S401_F_POST_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-48278	GSM1174122_MJH_STRRIDE_S317_F_POST_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-48278	GSM1174123_MJH_STRRIDE_S317_F_PRE_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-48278	GSM1174116_MJH_STRRIDE_S301_E_POST_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-48278	GSM1174096_MJH_STRRIDE_S235_C_POST_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-48278	GSM1174061_MJH_STRRIDE_S172_A_PRE_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-48278	GSM1174053_MJH_STRRIDE_S156_A_PRE_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-48278	GSM1174050_MJH_STRRIDE_S146_C_POST_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-48278	GSM1174106_MJH_STRRIDE_S257_D_POST_1_HG-U133_Plus_2_.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-62203	8 samples	Whole series due to Low percent present and other fluctuations
E-GEOD-7014	GSM161970.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-7014	GSM161944.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-7014	GSM161943.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-MTAB-37	A-673_SS271874_HG-U133_Plus_2_HCHP-167937_.CEL	RLE wider than +-0.2
E-MTAB-37	RD_SS275763_HG-U133_Plus_2_HCHP-170309_.CEL	RLE wider than +-0.2
E-GEOD-18732	GSM465386.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-18732	GSM465281.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-18732	GSM465319.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2

E-GEOD-9103	GSM230397.cel	RLE wider than +-0.2
E-GEOD-9103	GSM230407.cel	RLE wider than +-0.3
E-GEOD-9103	GSM230418.cel	RLE wider than +-0.4

Table S3 | Samples and series removed from the mouse microarray data collection because they failed quality controls. In total 122 samples were considered of low quality and were not used for any further analyses. All samples are removed from the gray shaded series.

Series	Samples	Reason
E-GEOD-12730	GSM319343.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13347	GSM313205.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-16438	GSM413181.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-16438	GSM413176.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-16438	GSM413161.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-16486	GSM414370.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-18033	56 samples	Whole series due to abnormal high percent present
E-GEOD-25908	GSM636278.cel	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-25908	GSM636225.cel	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-38870	18 samples	Whole series due to Low percent present and other fluctuations
E-GEOD-43373	GSM1061639_Mus_SE2__D4_13515.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-43373	GSM1061638_Mus_SE2__D4_13514.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-43779	GSM1071181_Rahme_04-06-10_2-AA_treated_muscle_4D_replicate_3.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-43779	GSM1071179_Rahme_04-06-10_2-AA_treated_muscle_4D_replicate_1.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-45577	GSM1109982_NUID-0000-0150-3235.cel	RLE off limits
E-GEOD-45577	GSM1109962_NUID-0000-0150-3224.cel	RLE off limits
E-GEOD-45577	GSM1109961_NUID-0000-0150-3205.cel	RLE off limits
E-GEOD-45577	GSM1109960_NUID-0000-0150-3200.cel	RLE off limits
E-GEOD-45577	GSM1109959_NUID-0000-0150-3238.cel	RLE off limits
E-GEOD-47104	GSM1144810_KT-13MB6D2F1old_5.14.08_2.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-47104	GSM1144808_KT-11MB6D2F1adult_5.14.08_2.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-6398	GSM147516.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-65927	GSM1611277_15_4semN1.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-65927	GSM1611276_14_4semB.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-65927	GSM1611275_13_4semA.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-7605	GSM183976.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-7605	GSM183977.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-MEXP-1623	C9.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE

		wider than +-0.2
E-MEXP-2446	681WTmuscleLADROSEMOU_03_080814.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-MEXP-2446	482WTmuscleShamROSEMOU_02_080814.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-12337	GSM309962.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13031	GSM326496.cel	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13874	GSM349106.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13874	GSM349107.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-13874	GSM349108.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-1479	GSM25168.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-19079	GSM472351.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-21368	GSM372908.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-23101	GSM569342.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-23101	GSM569339.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-30164	4 samples	Very low percent present 1.4-4%
E-GEOD-50399	GSM1218142_3wks_cko3.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-5500	GSM126911.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-62049	GSM1518961_CD117310DN_Mouse430_2_.CEL	RLE off limits
E-GEOD-7424	GSM179576.CEL	NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-8199	GSM202774.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2
E-GEOD-58676	GSM1416750_GFP_S48_3.CEL	Low percent present compared to other samples from the same series. Also NUSE above 1.1 and RLE wider than +-0.2

Table S4 | Number of samples, series and expressed genes for each of 69 and 73 categories in human and mouse respectively.

Organism	Category (Anatomic part, condition, gender, age, strain mouse specific)	Samples	Series	Expressed genes
Human	Heart, Normal, Both, All	60	10	5280
	Heart, Normal, Male, All	43	9	4934
	Heart, Normal, Female, All	17	6	5879
	Heart, Normal, Both, Old	15	4	4289
	Heart, Normal, Both, Adult	20	5	4535
	Heart - Myocardium, Normal, Both, All	19	4	3773
	Heart - Left ventricle, Normal, Both, All	31	6	6108
	Heart - Left ventricle, Normal, Female, All	12	5	6397
	Heart- Left ventricle, Normal, Male, All	19	5	5873
	Heart - Both ventricles, Normal, Both, All	37	6	5899
	Heart - Left ventricle, Idiopathic cardiomyopathy, Both, All	26	1	6184
	Heart - Left ventricle, Idiopathic cardiomyopathy, Male, All	16	1	6319
	Heart - Loth ventricles, Arrhythmogenic right ventricular cardiomyopathy, Both, All	12	1	3880
	Heart, Dilated cardiomyopathy, Both, All	35	2	4451
	Heart - Myocardium, Dilated cardiomyopathy, Both, Adult	21	1	3809
	Heart - Both ventricles, Dilated cardiomyopathy, Both, All	14	1	5111
	Heart, Ischemic cardiomyopathy, Both, All	55	6	6536

	Skeletal muscle - Rectus abdominis, Upper gastrointestinal cancer, Both, All	17	1	4035
	Skeletal muscle - Quadriceps, Type 2 diabetes DM2, Both, All	68	4	3738
	Skeletal muscle - Quadriceps, Type 2 diabetes DM2, Female, Adult	16	2	3268
	Skeletal muscle - Quadriceps, Type 2 diabetes DM2, Male, All	52	4	3913
	Skeletal muscle - Quadriceps, Thiazolidinedione TZD PPAR gamma ligand treatment for 3 months, All, All	16	1	3722
	Skeletal muscle - Quadriceps, Septic, All, All	12	1	4741
	Skeletal muscle - Quadriceps, Pre type 2 diabetes DM2, Male, All	12	1	5653
	Skeletal muscle - Quadriceps, Post hyperinsulinemic euglycemic clamp, All, All	16	1	4000
	Skeletal muscle - Quadriceps, Post hyperinsulinemic euglycemic clamp thiazolidinedione TZD PPAR gamma ligand treatment for 3 months, All, All	17	1	4065
	Skeletal muscle, Myotonic dystrophy type 2, All, All	20	1	3863
	Skeletal muscle - Quadriceps, Insulin resistant polycystic ovary syndrome PCOS, Female, Adult	16	1	2828
	Skeletal muscle - Quadriceps, Insulin resistant, All, Adult	38	1	3724
	Skeletal muscle - Quadriceps, Insulin resistant thiazolidinedione TZD PPAR gamma ligand treatment for 3 months, All, Adult	46	1	3721
	Skeletal muscle - Quadriceps, Insulin resistant post hyperinsulinemic euglycemic clamp, All, Adult	42	1	3955
	Skeletal muscle - Quadriceps, Insulin resistant post hyperinsulinemic euglycemic clamp thiazolidinedione TZD no response to treatment, All, Adult	12	1	3582
	Skeletal muscle - Quadriceps, Insulin resistant post hyperinsulinemic euglycemic clamp thiazolidinedione TZD response to treatment, All, Adult	12	1	3631
	Skeletal muscle - Quadriceps, Insulin resistant post hyperinsulinemic euglycemic clamp thiazolidinedione TZD unresponsive to treatment, All, Adult	25	1	3556
	Skeletal muscle - Quadriceps, Glucose intolerant, All, All	26	1	3408
	Skeletal muscle - Rectus abdominis, Coronary Artery Disease, All, All	61	1	4339
	Skeletal muscle - Quadriceps, Chronic Obstructive Pulmonary disease sedentary, All, Old	15	1	4580
	Skeletal muscle - Quadriceps, Chronic Obstructive Pulmonary disease trained, All, old	15	1	4407
	Skeletal muscle - Quadriceps, Chronic Obstructive Pulmonary disease, All, Old	30	1	4509
	Skeletal muscle - Quadriceps, Calorie restrictive for 12 weeks, Female, All	14	1	3484
	Skeletal muscle, Normal, All, All	1107	46	5257
	Skeletal muscle, Normal, Female, All	438	33	5524
	Skeletal muscle, Normal, Male, All	666	41	5103
	Skeletal muscle - Biceps brachii, Normal, All, All	45	5	3984
	Skeletal muscle - Quadriceps, Normal, All, All	994	32	5298
	Skeletal muscle - Rectus abdominis, Normal, All, All	13	2	4050
	Skeletal muscle - Quadriceps, Normal, Female, All	379	20	5631
	Skeletal muscle - Quadriceps, Normal, Male, All	614	28	5125
	Skeletal muscle, Normal, All, young	192	17	4071
	Skeletal muscle, Normal, All, adult	565	25	5298
	Skeletal muscle, Normal, All, old	261	21	7020
	Skeletal muscle - Quadriceps, First degree diabetes relative, All, All	39	2	3656
	Skeletal muscle - Quadriceps, Reported protein intake 0.75 g kg, Male, All	22	1	3115
	Skeletal muscle - Quadriceps, Reported protein intake 0.75 g kg, Male, adult	12	1	3154
	Skeletal muscle - Quadriceps, Reported protein intake 1.00 g kg, Male, All	22	1	3018
	Skeletal muscle - Quadriceps, Reported protein intake 1.00 g kg, Male, adult	12	1	3059
	Skeletal muscle - Quadriceps, Reported protein intake 0.50 g kg, Male, All	22	1	3178
	Skeletal muscle - Quadriceps, Reported protein intake 0.50 g kg, Male, adult	12	1	3188
	Skeletal muscle, Resistance exercise, All, All	114	6	6167
	Skeletal muscle, Resistance exercise, Female, All	42	6	7154
	Skeletal muscle, Resistance exercise, Male, All	73	6	5827
	Skeletal muscle, Resistance exercise, All, Young	39	4	4319

	Skeletal muscle - Quadriceps, Resistance exercise, All, Adult	45	3	7172
	Skeletal muscle - Quadriceps, Resistance exercise, All, Old	30	3	7817
	Skeletal muscle - Quadriceps, Trained, All, All	38	2	7915
	Skeletal muscle - Quadriceps, Endurance exercise, All, All	42	2	3170
	Skeletal muscle - Quadriceps, Aerobic exercise, All, All	27	1	5297
	Skeletal muscle - Quadriceps, Sedentary, All, All	38	3	3493
	Skeletal muscle, DMD, Male, Child	16	1	6400
Mouse	Heart, Normal, Both, All, All	296	65	5437
	Heart, Normal, Male, All, All	219	49	5371
	Heart, Normal, Female, All, All	63	19	5479
	Heart, Normal, Both, Young, All	161	38	5763
	Heart, Normal, Male, Young, All	107	26	5765
	Heart, Normal, Female, Young, All	44	13	5578
	Heart, Normal, Both, Adult, All	80	16	4834
	Heart, Normal, Both, Old, All	32	9	5508
	Heart, Normal, Male, Young, C3H-HeJ	16	1	6140
	Heart - Cardiomyocyte, Normal, Both, All, All	18	5	6178
	Heart - Left ventricle, Normal, Both, All, All	38	9	5302
	Heart - Both ventricles, Normal, Both, All, All	55	12	5569
	Heart, Normal, Both, All, CD1	28	4	6359
	Heart, Normal, Both, All, C57BL-6J	140	33	5318
	Heart, Normal, Male, All, C57BL-6J	105	27	5183
	Heart, Normal, Female, All, C57BL-6J	26	6	5441
	Heart, Normal, Both, Young, C57BL-6J	76	18	5261
	Heart, Normal, Female, Young, C57BL-6J	23	5	5380
	Heart, Normal, Male, Young, C57BL-6J	47	12	4829
	Heart, Normal, All, Adult, C57BL-6J	36	9	4802
	Heart, Normal, Male, Old, C57BL-6J	16	5	5493
	Heart, Normal, Both, Embryo, All	88	10	7318
	Heart, Aortic banding, Both, All, All	14	3	6538
	Heart, Calorie restricted diet, Both, All, All	15	3	4529
	Heart, Sham, Both, All, All	23	4	5566
	Heart, Transverse aortic constriction, Both, All, All	14	2	5765
	Skeletal muscle - Precursor cells, Normal, Male, All, All	14	1	4125
	Skeletal muscle - Gastrocnemius, Tenotomy, Male, Young, C57BL-6J	17	1	4439
	Skeletal muscle - Gastrocnemius, Sham, Both, Young, All	17	2	4638
	Skeletal muscle, Sham, Male, All, All	13	2	5552
	Skeletal muscle, Calpain3 knockout, Male, All, All	21	1	4745
	Skeletal muscle - Tibialis anterior, Cardiotoxin injection, Male, Adult, C57BL-6J	20	1	8245
	Skeletal muscle - Gastrocnemius, Casting, Male, Young, C57BL-6J	25	1	5885
	Skeletal muscle - Tibialis anterior, Glycerol injection, Male, Adult, C57BL-6J	18	1	8672
	Skeletal muscle, Mdx, Male, Young, All	16	4	6445
	Skeletal muscle, High fat diet, Both, All, All	99	6	4727
	Skeletal muscle, High fat diet, Male, Young, All	16	2	4692
	Skeletal muscle, High fat diet, Both, Adult, All	83	6	4714
	Skeletal muscle - Gastocnemius, High fat diet, Both, Adult, C57BL-6J	15	2	4603
	Skeletal muscle - Quadriceps, High fat diet, Male, Adult, C57BL-6J	22	2	4614

	Skeletal muscle, Normal, Both, All, All	346	57	5216
	Skeletal muscle, Normal, Female, All, All	44	17	4860
	Skeletal muscle, Normal, Female, Adult, All	14	4	4786
	Skeletal muscle, Normal, Female, Young, All	15	7	5270
	Skeletal muscle - Quadriceps, Normal, Female, All, All	15	5	4747
	Skeletal muscle, Normal, Both, Old, All	42	8	5117
	Skeletal muscle - Gastrocnemius, Normal, Male, Old, All	18	4	5803
	Skeletal muscle - Quadriceps, Normal, Male, Old, All	14	3	4591
	Skeletal muscle, Normal, Both, Adult, All	119	19	5033
	Skeletal muscle, Normal, Both, Adult, C57BL-6J	59	8	4884
	Skeletal muscle, Normal, Male, Adult, FVB	15	2	5128
	Skeletal muscle - Gastrocnemius, Normal, Both, Adult, All	32	6	4535
	Skeletal muscle - Quadriceps, Normal, Both, Adult, All	49	6	4996
	Skeletal muscle, Normal, Both, Young, All	145	26	5381
	Skeletal muscle, Normal, Both, Young, C57BL-6J	79	11	5473
	Skeletal muscle - Quadriceps, Normal, Both, Young, All	15	5	4492
	Skeletal muscle, Normal, Male, All, All	305	46	5294
	Skeletal muscle - Gastrocnemius, Normal, Both, All, All	127	16	5473
	Skeletal muscle - Quadriceps, Normal, Both, All, All	81	14	4892
	Skeletal muscle - Soleus, Normal, Both, All, All	22	3	4884
	Skeletal muscle - Tibialis anterior, Normal, Both, All, All	15	5	6971
	Skeletal muscle, Normal, Male, Old, All	34	6	5172
	Skeletal muscle, Normal, Male, Adult, All	105	15	5056
	Skeletal muscle, Normal, Male, Adult, C57BL-6J	52	6	4930
	Skeletal muscle - Gastrocnemius, Normal, Male, Adult, All	23	4	4646
	Skeletal muscle - Quadriceps, Normal, Male, Adult, All	45	5	4925
	Skeletal muscle, Normal, Male, Young, All	127	21	5403
	Skeletal muscle, Normal, Male, Young, C57BL-6J	72	8	5506
	Skeletal muscle - Gastrocnemius, Normal, Male, Young, All	76	8	5665
	Skeletal muscle - C2C12, Normal undifferentiated, Female, All, All	20	6	7121
	Skeletal muscle - C2C12, Normal 1-2 days differentiated, Female, All, All	39	2	5815
	Skeletal muscle - C2C12, Normal 3-4 days differentiated, Female, All, All	14	5	6999
	Skeletal muscle - C2C12, Normal 5 days differentiated, Female, All, All	12	4	7366

Table S5 | Samples that were predicted to have opposite gender from what was reported on the metadata but turned out to be copying errors.

Organism	Series	Samples	Reported gender	Prediction | Reason
Human	GSE13205	GSM333436	60 years old septic Male	Female | The corresponding publication reports only a 60 years old female
	GSE3526	GSM80654, GSM80658, GSM80790	All females	All males | These samples were identified as duplicates. The original IDs report them as males
	GSE38780	GSM949391	17 years old	Male | The publication states one 17

			female	years old male
Mouse	E-MEXP-733	All samples	Mixed gender	All samples were strongly predicted as opposite gender | Possible copying error
	GSE25729	GSM632001	Male	Female | Possible copying error
	GSE1479	Samples past E11	All females	Mixed gender | Gender differentiation in mice happens between E11 and E12 from which we had mixed gender predicted

S1 | A two step PHP script that first calculates and assigns MD5 hash keys to all files within a folder and then compares them to identify the duplicate files. Useful to find duplicate raw CEL files with different IDs.

```php
<?php
// ------------------------------------------------------------------------
// A fast and simple script that finds the MD5 hash keys
// for all files within a folder and compares them
// in order to detect duplicates.
//
// ------------------------------------------------------------------------
// How to execute this script from the command line interface (CLI):
// "path/to/php" "path/to/this_script.php" "path/to/CEL_directory/"
//
// ------------------------------------------------------------------------
// If you do not have PHP (CLI) installed in your computer, you can
// download the latest version from http://php.net/downloads.php
//
// ------------------------------------------------------------------------
// Author: Apostolos Malatras, email: apmalatras@biol.uoa.gr
// Date: April 21, 2016
// ------------------------------------------------------------------------

$argument_1 = $argv[1];

if(is_dir($argument_1)){
	$dir=scandir($argument_1);
	$dirnum=count($dir);

	//calculate hash keys
	for($i=2;$i<$dirnum;$i++){
		$dir[$i]=trim($dir[$i]);
		//Hash array
		$md5array[$dir[$i]] = md5_file("$argument_1/$dir[$i]");
	}
	//compare
```

```
        $dup=array();
        $rep=array();
        foreach(array_count_values($md5array) as $val => $c){
                if($c > 1){
                        $dup[] = $val;
                }
        }
        foreach($dup as $key_dup => $val_dup){
                foreach($md5array as $key_md5 => $val_md5){
                        if($val_dup == $val_md5){
                                $rep[$val_dup].= "$key_md5\t";
                        }
                }
                $rep[$val_dup] = trim($rep[$val_dup]);
        }

        if($rep==NULL){
                echo "No duplicate files detected\n";
        }else{
                print_r($rep);
        }
}else{
        die("First argument must be a directory\n");
}
?>
```

2.3 "Annexin A2 links poor myofiber repair with inflammation and adipogenic replacement of the injured muscle"

Published in: Human Molecular Genetics, Volume 26, Issue 11, 1 June 2017, Pages 1979–1991

Link: https://doi.org/10.1093/hmg/ddx065

2.4 "CellWhere: graphical display of interaction networks organized on subcellular localizations"

Published in: Nucleic Acids Research, Volume 43, Issue W1, 1 July 2015, Pages W571–W575

Link: https://doi.org/10.1093/nar/gkv354

Chapter 3 – Discussion

In this thesis we developed two muscle-specific bioinformatics tools, CellWhere and Myominer, which will help myologists and other researchers in the analysis and interpretation of various tissue-, cell- or pathology-specific elements of gene expression, regulation, function and protein localizations and interactions. We also accumulated and analyzed a substantial proportion of all publicly available muscle-related microarray data that could also be used in further tools or studies.

3.1 CellWhere tool

With CellWhere users can input a list of genes and discover protein localization from Uniprot, GO or both. The localizations can be prioritized on their annotation frequency or by premade priority scores (flavors): muscle, secretory and mitochondria. Custom-made priority flavours are also supported, allowing researchers to adapt CellWhere to their field of interest. The interactive network resembles the cell and proteins (nodes) placed on appropriate compartments. Edge-thickness depicts the interaction score and can be clicked to reveal its value alongside the relevant publications that it is derived from (evidence). Nodes and intermediate subcellular compartments are interactive and can be moved to different locations in order to create a visually clear network for publication or sharing. CellWhere automatically retrieves the non-redundant manually-annotated compressed version of Uniprot (Swiss-Prot) automatically within 24 hours of its monthly release. The same automated process is used weekly to acquire protein interactions and scores from the Mentha interactome browser (Calderone, Castagnoli et al. 2013). Identifiers, UniProt/GO localizations and interactions are parsed and stored in a relational database. The UniProt and GO localizations are collapsed to 50 CellWhere localizations which include all major cell compartments. Proteins that form many non-pathway-specific interactions (e.g. ubiquitins), are removed them from the final output, using a filter. By default the filter removes proteins that bind more than 100 partners, corresponding to 1.6% of the total number of proteins. All data are stored on the Openshift cloud in two separate virtual

machines: one used for update, raw compressed data parsing and analysis and the second for data storage and the web interface. CellWhere summarizes subcellular localizations and local interactions quickly and accurately while visualizing them on an interactive network.

CellWhere has already been cited by several researchers who have used it to map proteins to subcellular localizations (He, Vanlandewijck et al. 2016; Simon, Murchison et al. 2017). It was also used for the representation of the most common cellular compartments of the dystrophin interactome project (https://sys-myo.rhcloud.com/dystrophin-interactome/index.html) (Thorley 2016).

After the completion of CellWhere, we accumulated the majority of published muscle pertinent raw transcriptomic microarray data. After researching preprocessing, quality control and analysis options, we configured an analysis pipeline for combinatory studies and a pipeline for differential expression analyses. The collected microarray data were already used for two studies (see Appendix) and for the creation of a muscle specific co-expression tool, MyoMiner.

3.2 MyoMiner database

Since the dramatic expansion and accumulation of gene expression data, pooled data analyses, such as co-expression or meta-analyses, could provide a better understanding of biological systems. In co-expression analyses, if gene expression levels are calculated on combined data from multiple experiments, higher statistical power can be achieved and interesting conclusions about multiple conditions can be drawn. Also the difference in co-expression between conditions can reveal potential gene regulators. The idea behind co-expression analyses is to determine gene function as genes that are correlated, in multiple samples, are likely to be involved in similar functions (guilt by association). As with other omics approaches, co-expression can be used to generate hypotheses for gene function and regulation.

In order to ensure high quality of the resulting databses, a large part of this thesis involved the curation and meticulous analysis, either with custom-written

programmatic scripts or manually, of the collected muscle microarray data. Thus, we tested the available high-throughput quality controls, pre-processing and downstream analysis methods, ranging from the initial ones introduced in the 90s to the newest. Moreover, we sought to gather any missing information from the metadata. We crosschecked the entries from the original publications, supplementary data or by communicating with the corresponding authors. Since we had a large collection of samples we also used algorithms or other databases to predict some of the missing values. For example, because half of our samples had their gender missing, we mapped the genes from each sample to their corresponding chromosomes and then classified the samples by gender, based on Y chromosome gene expression. This method had more than 98% accuracy and we were even able to detect and correct 21 samples that were present on the GEO and AE repositories (see MyoMiner "Data statistics" section and table S5). We sent our findings to GEO and AE curators and most of them have been corrected.

For MyoMiner we selected the microarray platforms from GEO and AE that are linked to the largest number of experiments for humans and mice. We processed the raw data using various methods which are now streamlined for easier insertion of new transcriptomic data in MyoMiner.

We also built a simple and easy-to-use web interface to search for the transcriptional co-expression of any expressed gene pair in muscle cells and tissues in various conditions. So far we have included 142 human and mouse categories based on age, gender, anatomic part and condition. Users select category and gene of interest and MyoMiner returns all expressed correlated gene-pairs with their r and adjusted p value. Follow up tools are included to narrow down the list of genes that may be functionally associated with initial gene, such as a standardized expression level scatterplot, a network creation tool and a comparison tool to search for differentially co-expressed genes. The calculated co-expression data are stored in the Okeanos and Openshift clouds. These co-expression analyses will help muscle researchers to delineate the tissue-, cell-, and pathology-specific elements of muscle protein

interactions, cell signaling and gene regulation. Changes in co-expression between pathologic and healthy tissue may suggest new disease mechanisms and therapeutic targets. MyoMiner is a powerful muscle-specific tool for the discovery of genes that are associated in related functions based on their co-expression.

MyoMiner was used in two analyses regarding dysferlin co-expression in normal muscle tissue: first to identify genes that are co-expressed with Dysferlin; and secondly to test how the gene co-expression of dysferlin's known protein binding partners changes across different muscle conditions.

In the first analysis, we accumulated the *DYSF* co-expressed genes with r values higher than 0.4 in all normal muscle tissues and muscle cells (11 categories) that are currently present on MyoMiner, in order to obtain an overview across the normal muscle (the categories are: human quadriceps, rectus abdominis, biceps brachii, heart, mouse quadriceps, tibialis anterior, gastrocnemius, soleus, heart and two C2C12 myotubes). We kept the genes that are present in at least 6 categories (Table 6). Several of these genes, including *OBSCN*, *ITGA7*, *FLNC, and CACNA1S,* are related with myo- and cardiomyopathies. Enrichment analysis of the co-expressed genes on Mouse Genome Informatics phenotypes returns the following terms: centrally nucleated skeletal muscle fibers, abnormal skeletal muscle fiber morphology, decreased skeletal muscle mass, abnormal sarcoplasmic reticulum morphology, myopathy, among others (Figure 16, left). Enrichment of the same genes on the Reactome pathway database (Fabregat, Sidiropoulos et al. 2016) outputs membrane trafficking and vesicle-mediated transport (Figure 16, right)

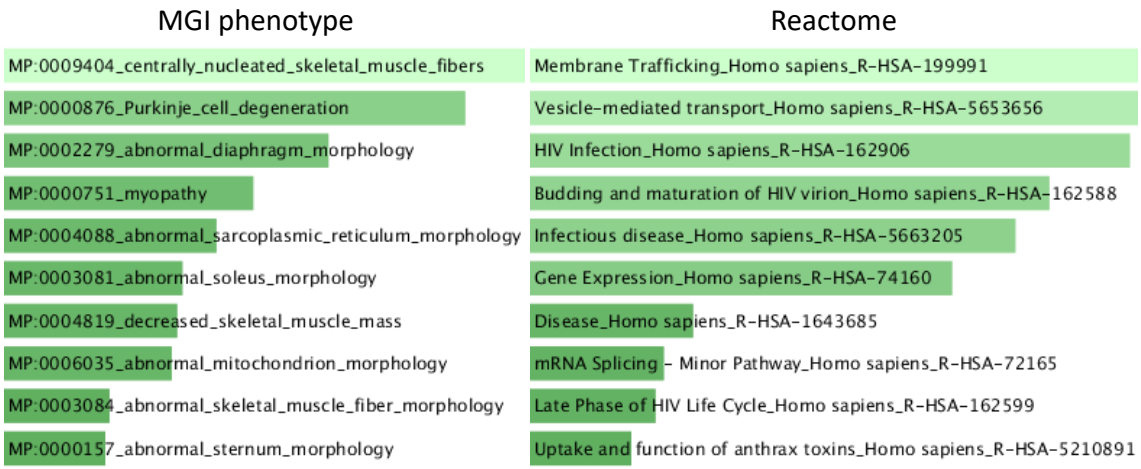

Figure 16 | Enrichment analysis of *DYSF* consensus co-expressed genes. On the left is the output of the enrichment analysis from Mouse Genome Informatics database of the DYSF consensus co-expressed genes. The genes that contribute to muscle related terms are *AFG3L2, SRPK3, OBSCN, ITGA7, FLNC* and *CACNA1S*. On the right is the output from Reactome pathway database. The genes associated with the first two results, membrane trafficking and vesicle-mediated transport are: *ARFRP1, MYO1C, TBC1D20, VPS4A, PPP6R3, VPS37C, AP1B1, KIF1C, SEC24C* and *AGPAT3*.

To understand the relationship of these co-expressed genes to existing knowledge of Dysferlin interactions, color-coding is used in table 6 to indicate *DYSF* interactors that are present in other databases: grey indicates interactors from PSICQUIC (Aranda, Blankenburg et al. 2011), blue are from (Assadi, Schindler et al. 2008) using tandem affinity purification mass spectrometry, orange are differentially expressed proteins from Bla/j mouse quadriceps using liquid chromatography tandem-mass spectrometry (LC-MC/MC) and green are differentially expressed genes from microarray studies. However, many of the co-expressed genes have not been characterized for their possible interaction with DYSF and may be important to understand the molecular mechanisms underlying dysferlinopathies.

EIF3B (10)	PPM1G (7)	CYHR1	NSUN2	SON
ITGA7 (10)	SEC24C (7)	D17WSU92E	NUDCD3	SRPK3
DHX16 (8)	TSC2 (7)	DENND4B	OBSCN	STAU1
NPLOC4 (8)	UBAC2 (7)	DUSP27	PACSIN3	TBC1D20
PPME1 (8)	USP7 (7)	ESYT1	PARP1	TCEB3
TTC7B (8)	UTP6 (7)	FLII	PDCD6IP	THUMPD1
ABCF1 (7)	VPS37C (7)	GYS1	PDCD7	TRAK1
ACTR1B (7)	VPS4A (7)	HIVEP2	PPP6R3	TTC17

AGPAT3 (7)	AARS	ICMT	PRPF8	UBE4B
ARFRP1 (7)	ABHD12	KIF1C	QRICH1	UBQLN4
CD99L2 (7)	AFG3L2	KPNB1	RRN3	USP22
CORO6 (7)	AGO2	MAP3K4	RRP12	USP5
EIF4E2 (7)	AI464131	MAP4	SAE1	ZDHHC7
FLNC (7)	AP1B1	MRPS27	SCAMP2	ZFYVE26
LRRC47 (7)	ATP6V1B2	MYO10	SCAMP3	
MAP2K4 (7)	CACNA1S	MYO1C	SEC14L1	
NEURL4 (7)	CDK16	NOMO1	SF3B3	
NXN (7)	CIPC	NRBP1	SH2B1	

Table 6 | Genes highly correlated with *DYSF* across several normal categories. Nine normal muscle tissues (human and mouse quadriceps and heart, mouse gastrocnemius, tibialis anterior, soleus, human biceps brachii, rectus abdominis) and two C2C12 myotube cells categories (3-4 and 5+ days after differentiation) were used to acquire the above list of co-expressed genes that have r value above 0.4 in at least 6 categories of the aforementioned categories (the number of categories is next to the gene in case of more than 6).

In the second analysis, to explore the gene co-expression of dysferlin with its known protein binding partners across different muscle conditions, we collected and mapped to gene symbols the dysferlin protein interactions using PSICQUIC view from EBI (Aranda, Blankenburg et al. 2011). We also concatenated the Spearman correlation r values of the interactors from different tissues and conditions and clustered them as shown (Figure 17). We used two C2C12 immortalized muscle cell categories, two muscle regeneration categories (cardiotoxin and glycerol), six skeletal muscle tissues and the heart in order to have an overview across muscle tissue and cell and conditions. Dysferlin gene expression correlates well with known protein binding partners in mouse and cell samples, but not in human samples. This may suggest that some of what is known from normal cell and mouse studies, since these are the origin of most protein binding data, is questionable in the human context and that more human tissue and cells transcriptomic studies are needed in the dysferlinopathic state. For example, *PARVB* is highly correlated on mouse samples and cells and much lower on human. It is known from previous studies (Cagliani, Magri et al. 2005) that in dysferlinopathic muscles, ANXA1 and ANXA2 gene and protein expression have an inverse relationship to DYSF expression, and in these muscle conditions we see this reflected in normal C2C12

cell samples and also in damaged muscle tissue (cardiotoxin and glycerol), where Annexin genes are highly anti-correlated with *DYSF* (Figure 17 top), but not in undamaged whole muscle. Several of the consistently co-expressed genes are related with neuromuscular disorders, including filamin C (*FLNC*) an actin-cross-linking protein, caveolin 3 (*CAV3*), desmin (*DES*), ring finger protein 10 (*RNF10*) and kinesin family member 1B (*KIF1B*).

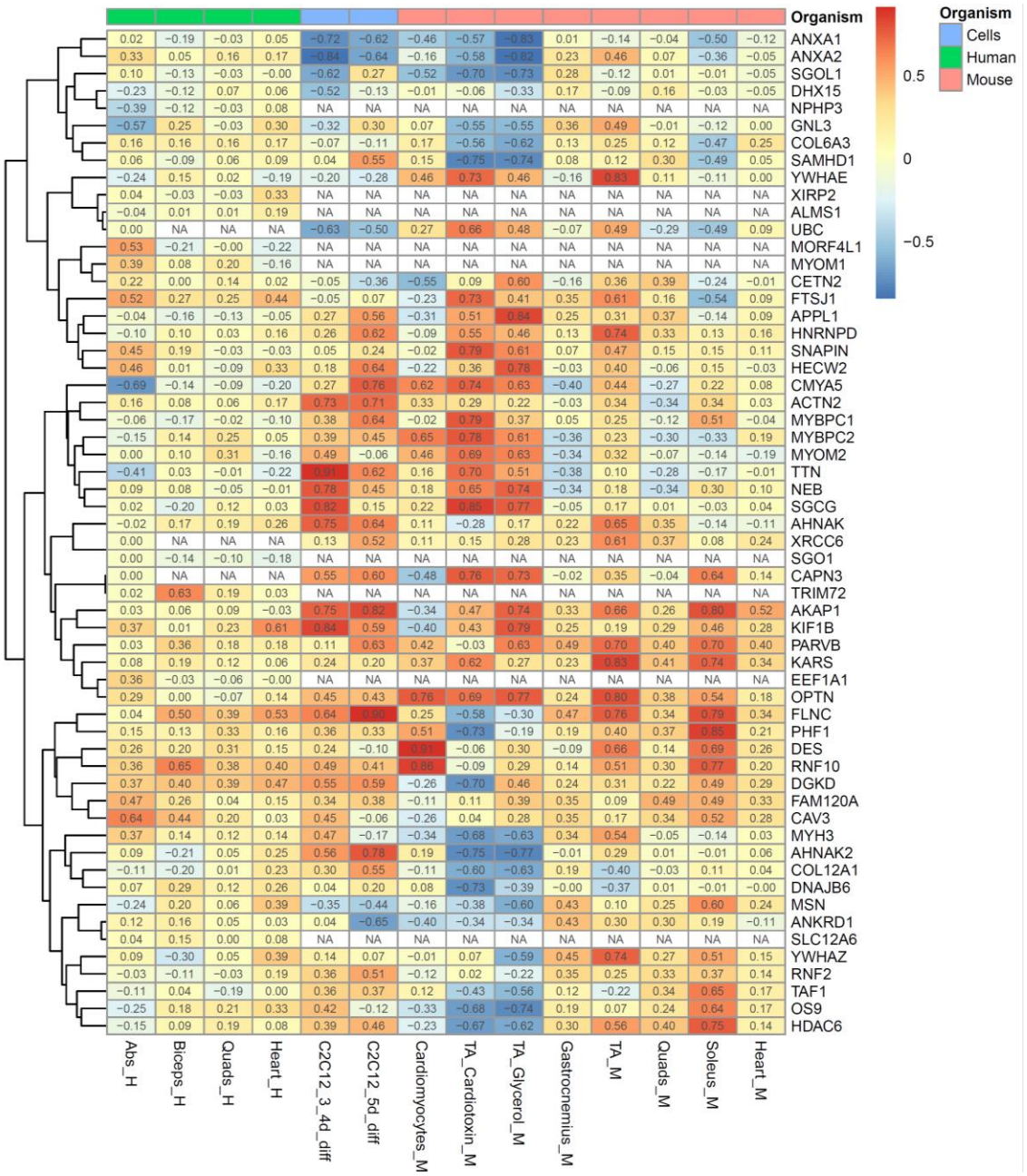

Figure 17 | Dysferlin interactors clustered with several co-expression categories. On the right are *DYSF* interactors acquired from PSICQUIC and their correlation values with *DYSF*. The categories include both genders and all ages unless otherwise stated: human rectus abdominis, biceps brachii, quadriceps, heart, C2C12 cells 3 to 4 days after differentiation, C2C12 cells 5 days and more after differentiation, mouse cardiomyocytes, tibialis anterior from adult male mice after cardiotoxin injection, tibialis anterior from adult male mice after glycerol injection, mouse tibialis anterior, quadriceps, soleus and heart. The genes that could not be measured by the microarrays have white background (NA).

The carefully pre-processed microarray data that we have accumulated have many more uses outside of the scope of MyoMiner. One application is to create collections (sets) consisting of the most differentially expressed genes in various normal or pathological muscle tissues or cells experiments. To create these muscle-specific gene sets (https://sys-myo.rhcloud.com/muscle_gene_sets.php - the new collection will be available on the website soon) we performed differential expression analysis on each series with our already established microarray analysis pipeline. Gene sets can be used for further downstream analyses and especially for gene set enrichment analysis (GSEA). Our "SysMyo" muscle gene sets have already been incorporated into the popular online enrichment tools Enrichr (Kuleshov, Jones et al. 2016) and WebGestalt (Wang, Vasaikar et al. 2017). We have also constructed gene set networks by using the Python implementation of Sets2Networks algorithm (Clark, Dannenfelser et al. 2012) on all the muscle gene sets, while retaining only the relations with higher than 20 % probability of interaction (~6,000 edges, the probability being used as a score) for the final network. From this network we selected their top connections for *DYSF* and *DMD* and constructed a predicted gene association network (list). We have also created specific networks for the *DYSF* and *DMD* neuromuscular disorders using only their related gene sets (>98 % interaction probability, ~500 edges). The predicted networks can suggest novel gene relationships or protein-protein interactors. For example, the *DYSF* gene association network has 9 genes in common with its protein-protein interaction network from PSICQUIC: *TTN, MYH3, CAV3, SGCG, MYOM2, MYOM1, FLNC, ACTN2, and NEB.* These networks are available at: https://sys-myo.rhcloud.com/MuscleGeneSets_networks

3.3 Dysferlin transcriptomic analyses

We also brought together dysferlinopathy microarray expression datasets and analyzed them in the context of Annexin-A2 (ANXA2) knockout (KO). We tested for similarities between three dysferlinopathic mice series (GSE2507, GSE2112 and GSE2629), ANXA2 KO microarrays and dysferlinopathy gene sets derived from older

studies (https://sys-myo.rhcloud.com/muscle_gene_sets.php). The dysferlin deficient microarray series showed significant overlap with dysferlin gene sets but the ANXA2 KO arrays did not show any overlap with them. We also tested for enrichment on inflammation-related processes. Seven related gene ontology terms where enriched in all microarray datasets but none of them (and the majority of the genes involved) were enriched on ANXA2 KO data. We also observed dysregulation of fat marker and fatty acid metabolism gene ontology terms in dysferlin deficient dataset. We analyzed another dataset comparing fat with muscle tissue to help interpret the results and found that fat-related terms where rather similar and genes where regulated with the same direction.

Our analysis of dysferlinopathic microarrays was stringent and thorough. For example, GSE2507 series is comprised of two experiments: SJL/J dysferlin deficient vs. C57BL/6 normal mice skeletal and cardiac muscles. From the skeletal muscles, n=5 samples are dysferlinopathic and n=5 are normal. A first look at the samples showed a good separation between the two conditions, but a clustering on the scan dates was also observed (Figure 18).

Sample	Scan date	Batch	Condition
GSM46161.CEL	11/04/03 09:30:07	2	DYSF - SM
GSM46162.CEL	11/13/03 08:57:16	3	DYSF - SM
GSM46163.CEL	11/13/03 09:13:04	3	DYSF - SM
GSM46164.CEL	11/13/03 09:27:19	3	DYSF - SM
GSM46165.CEL	11/04/03 09:41:41	2	DYSF - SM
GSM46166.CEL	10/30/03 09:14:45	1	WT - SM
GSM46167.CEL	10/30/03 09:27:09	1	WT - SM
GSM46168.CEL	10/30/03 09:51:20	1	WT - SM
GSM46169.CEL	11/13/03 09:41:27	3	WT - SM
GSM46170.CEL	10/30/03 10:04:07	1	WT - SM

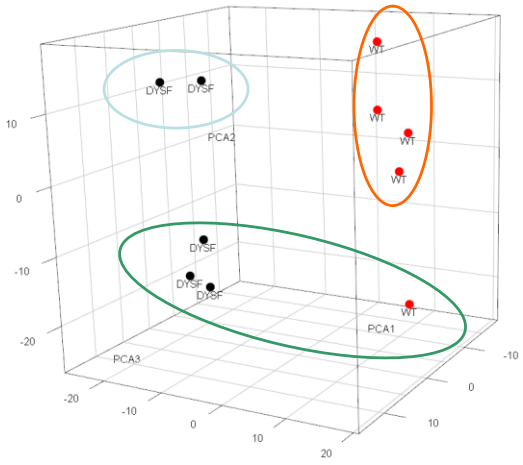

Figure 18 | Sample clustering in dysferlinopathy study GSE2507. On the left is a table with the skeletal muscle samples from GSE2507 alongside their scan dates. We separated them in batches based on the different scan dates. On the right is a PCA plot showing a good separation between dysferlin deficient and normal samples. However, it can also be seen that samples that were scanned on the same day are clustered

together. This could have led to technical artifacts in the results of the analysis, so we corrected it during data processing.

Further quality controls showed a low percent present of the GSM46169 sample 31.87% compared to the ~40% average, so this sample was removed (Figure 19 left). We then estimated the proportion of variation with the Principal Variable Components Analysis (PVCA) (Boedigheimer, Wolfinger et al. 2008) method for the scan dates and the biological group (DYSF vs WT) (Figure 19 A).

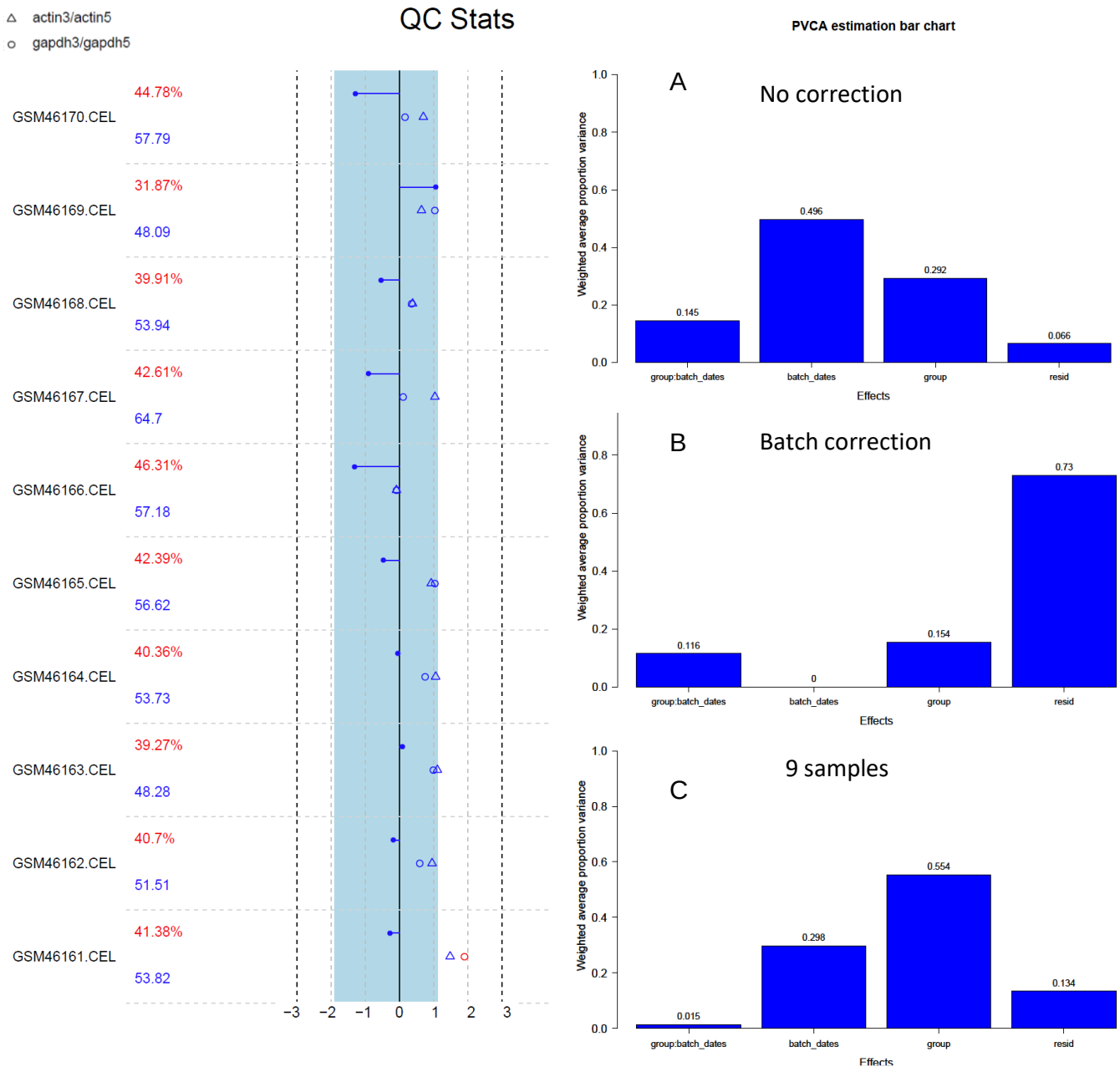

Figure 19 | GSE2507 skeletal muscle quality control and batch effects signal estimation. On the left is a typical Affymetrix quality control output. We see the percentage of present (expressed genes) next to each sample ID. In this case one of the samples, GSM46169, is colored red as it has more than 10% of expressed gene variation with the other samples. Below the percentage of present genes is the background noise value. The β-actin 3' to 5' ration is shown with a triangle and GAPDH ratio with a circle. Also the scaling factor is shown as the blue lines with a filled circle at their end, which should be inside the light blue background. On the right we see the proportion of batch effects using the date as a source. The first bar (group:batch_dates) shows the combined biological and date variation. The second (batch_dates) shows the variation

derived (technical variation) from the different scan dates. In this case (**A**) we see that most of the variation (and perhaps most of the differential expression) is due to the different dates that the microarrays were scanned. The third column (group) shows the biological and the fourth the remaining unknown variation. When the samples are batch corrected (**B**) we see that the scan date variation (batch_dates) is reduced to zero, although the biological (group) variation is also reduced. The bottom barplot (**C**) shows the variation when we removed the GSM46169 sample which has much less percent present genes. The biological variation is greatly increased to 55.4%, while the scanned date variation is reduced.

Since the scanned date accounts for almost half (49.6%) (Figure 19 A) of the differentially expressed genes, we used the ComBat and SVA (Leek, Johnson et al. 2012) algorithms to reduce the technical variation. Even though technical variation was completely removed, biological variation was also reduced (Figure 19 B). Thus, we concluded to only remove the low quality GSM46169 sample, which increased the biological variation while reducing the scan date variation (Figure 19 C). After pre-processing with the RMA algorithm and BrainArray CDF, we performed differential expression analysis with the Characteristic Direction algorithm (Clark, Hu et al. 2014) and continued with enrichment analysis using Enrichr (Kuleshov, Jones et al. 2016). This careful re-analysis yielded enrichment results that were much more muscle-relevant than those of the original publication (Figure 20).

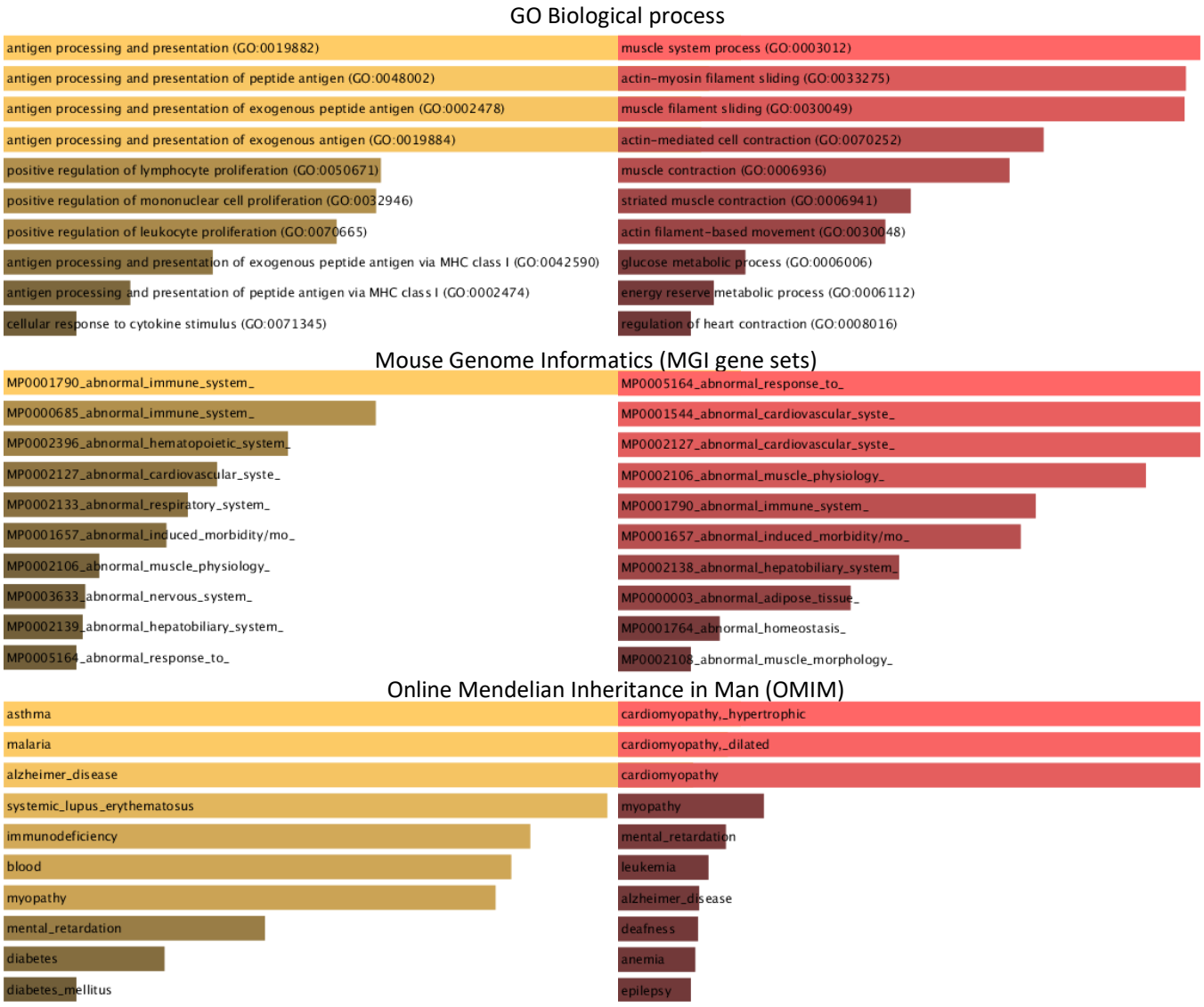

Figure 20 | Enrichment analysis of differentially-expressed genes in study GSE2507, comparing original and state-of-the-art methods. On the left we used the top 250 differentially expressed genes from the original publication for enrichment analysis on three databases using Enrichr: Gene Ontology biological process, Mouse Genome Informatics (MGI) and Online Mendelian Inheritance in Man (OMIM). On the right we used the top 250 differentially expressed genes for enrichment analysis, using our stringent pipeline. In all three enrichment analyses we see much more relevant results.

3.4 Co-expression on high throughput genomic data

Gene co-expression measures how similar the expression levels of different genes are, under the same or different conditions. Several analysis steps on raw data

are required to obtain to this information. Therefore, several co-expression databases have been constructed to provide easy access to co-expression information alongside other bioinformatics tools. Below we discuss some of the decisions and caveats when constructing co-expression databases.

Co-expression values (between -1 and +1) are calculated to show how similar is the expression of two genes under a certain condition is. Selecting the data to define a condition can be done in two ways. In a condition-dependant approach (Aoki, Ogata et al. 2007) the datasets are divided into categories and analyzed separately. Categories can be different tissues, developmental stage, phenotypes, etc. The co-expression values can vary, based on different conditions and with this approach the r-values can be used to calculate the difference in co-expression per condition. In a condition-independent approach (Aoki, Ogata et al. 2007) the goal is to use as many different conditions as possible. This analysis can show underlying gene pair relationships, independent of phenotypes or tissue type. This method is more appropriate for an initial screening of gene pair relationships.

Proper normalization of the data prior to correlation is required, since some correlation coefficients are sensitive to outliers. Most of the co-expression databases use the Affymetrix GeneChip technology as they are the most abundant microarrays. The popular pre-processing algorithms are MAS 5.0, RMA, GC-RMA. Only MAS 5.0 is using the information from MM probes and normalizes each sample independently. Although RMA is claimed to be superior to MAS 5.0 for differential expression (Jiang, Leach et al. 2008), it is not clear that this hold also for co-expression analysis. Lim *et al.* (Lim, Wang et al. 2007) concluded that MAS 5.0 would be the best option in this case, because inter-array pre-processing (RMA, GC-RMA, PLIER) introduces artificial correlations. However, MAS 5.0 computes expression values on a linear scale (and usually returns values below 1) by default and must be log-transformed to approximate normal distribution before using parametric correlation (e.g. Pearson). A better single-array normalization alternative is the SCAN algorithm (Piccolo, Sun et al. 2012) which

also corrects GC bias and reduces probe and array variation from each individual sample.

The choice of correlation coefficient also plays a major role. Most of the publicly available co-expression databases use Pearson correlation. It ranges from -1, meaning that genes tend to respond in opposite directions (anti-correlated) to +1, where genes respond the same way in all samples. Zero correlation represents no association. A drawback of Pearson correlation is its sensitivity to outliers, as one low quality sample could drive a false relationship. A robust alternative is Spearman correlation which can be obtained by ranking the gene expressions across samples before using the Pearson correlation formula. However, sometimes outliers could have biological meaning. Thus screening the expression values is important. Spearman correlation measures monotonic relationship, in contrast to Pearson that measures linear correlation. Monotonic correlations occur more frequently than linear ones. There are also other kinds of relationships like the mutual information (MI) (information theory) which has been used to identify relationships between genes (Steuer, Kurths et al. 2002). It assumes zero value in case of independence and unlike correlation, it has no upper limit for its relation score. However, because of the way MI is calculated, many more samples are required for the calculation of its score than the estimation of correlation coefficients. One of the most popular network construction tools ARACNE uses MI exclusively (Margolin, Nemenman et al. 2006).

One caveat of gene expression correlation is that it can be driven by other factors. For example, a transcription factor (TF) when is upregulated, drives the expression of gene X and Y. In this scenario, TF with X and TF with Y will be highly correlated. However, X and Y will be highly correlated as well, since both are upregulated from the same TF. This can be beneficial as X and Y could be involved in the same processes, but if we are interested specifically in the relation of X with Y, their correlation would be zero if TF was not upregulated. In order to extract the correlation between X and Y without TF interfering, we should calculate the partial correlation (Yule 1907). Partial correlation could theoretically be used to remove all the gene effects from

a pair of genes, but it would require more microarray experiments than the genes. It has been used successfully to create relatively small networks (Ma, Gong et al. 2007).

The q-value of a co-expression can be calculated by transforming the r-values to scores that approximately follow a *t* distribution and then adjusting for multiple testing by controlling the false discovery rate either with Bonferroni or the less conservative BH method. In cases where many samples were used to calculate gene co-expression, even with the multiple hypothesis adjustment, r-values as low as 0.2 can be significant and the abstract q-value cut-off of 0.05 will not be of practical importance. Therefore, on categories with many samples we recommend using lower cut-offs: at least 0.005 or even lower. Another way of determining co-expression significance is to use the q-value in conjunction with the confidence intervals (see MyoMiner manuscript) or the coefficient of determination r^2. Coefficient of determination is simply the r-value squared and it measures the scale of shared variance between the genes in question (from 0 - no shared variance to 1 – 100 % shared variance).

3.5 Training k-nearest-neighbor (k-NN) classifiers to predict specific muscle tissues

One of the direct uses of the microarray data collected in MyoMiner is to use them to train a tissue classifier. Specifically, we trained a k-nearest-neighbor (k-NN) classifier in order to predict specific skeletal and cardiac muscles. The classifier could help distinguish muscle anatomic parts when they are not given in the metadata or the original publications. For example, in many experiments the tissue is specified as skeletal muscle or heart but the exact part (e.g. quadriceps femoris or vastus lateralis, etc) is not provided. We constructed two classifiers for human: one for the skeletal muscles and one for the cardiac muscles. Even though both groups are categorized as striated muscles their genetic profile is quite different. We also constructed the corresponding classifiers for mouse.

First we trained (train/test split was 80/20) the classifier without any dimensionality reduction methods and got poor predictions (accuracy ~ 0.7). Then we

applied multidimensional scaling (MDS also known as principal coordinates analysis) (Gower 1966) to reduce the data dimensions (genes) to 100 or 50, or even to first select the highly expressed genes using UPC percentages (Piccolo, Withers et al. 2013) and then reduce them to 100 or 50. We then trained the k-NN with a repeated n-fold cross validation with 10 folds and 15 repeats (Table 7). Even though the accuracy of the classifiers is rather high, when tested on samples outside of the training and testing set the results were mixed. When an experiment contained samples from a class that appeared in the training set, the classifier predicted the majority of samples correctly. We could have then infer that the erroneous predicted samples fall in the same category as the correctly predicted, because most of the times, researchers, gather samples from the same tissue (e.g. if all samples are predicted as quadriceps and one as vastus lateralis, we can say that this was incorrectly predicted as vastus lateralis as researchers usually take their samples from the same tissues). However, if the data were from an unknown class, all the predictions were wrong so we did not use these classifiers to predict the missing anatomic parts in our data thus far.

Classifier	Best k	Accuracy	Classes
Human cardiac muscles	7	0.9540	**4:** atrium, left ventricle, right ventricle, myocardium
Human skeletal muscles	3	0.9719	**7:** biceps brachii, deltoid, extraocular, paravertebral, quadriceps, rectus abdominus, vastus lateralis
Mouse cardiac muscles	5	0.9615	**4:** atrial, cardiomyocytes, myocardium, ventricle
Mouse skeletal muscles	5	0.9858	**4:** gastrocnemius, quadriceps, soleus, tibialis anterior

Table 7 | k-NN classifiers for specific muscle tissues. The accuracy of the classifiers is quite high. However, since we wanted to predict muscle tissues that could potentially belong to other classes, we did not use the classifiers for any further predictions.

3.6 Microarray limitations

Microarrays have been extremely useful in a wide area of biological applications but they also have a number of limitations. Most importantly, a microarray can only detect RNA sequences that the designed probes can detect. Simply put, if the RNA contains sequences that have no corresponding oligos in the array, the sequences will not be measured. In gene expression analysis, a gene that was not described before will not be present in the array. Also non-coding RNA sequences are typically not present on arrays. This problem is more pronounced in older arrays where only a set number of probes could be printed on the array; thus a portion of the genes could eventually be measured (e.g. Affymetrix Murine Genome U74Av2). Newer commercial arrays have tried to compensate for this by including probes that do not match to any known genes at the time they are designed - transcripts which can then be assigned to newly discovered genes if their sequences match. Also, as time progresses more researches are using the now popular BrainArray CDF (Dai, Wang et al. 2005), which is updated with new information annually.

Another difficulty in terms of probe design, is to generate probes of which the RNA sequences do not overlap. If sequences are homologous, then a probe could detect multiple genes at once, which is particularly problematic for genes with many splice variants or for genes that belong to the same family. Dai *et al.* (Dai, Wang et al. 2005), address this issue by selecting probes that detect specific and unique parts of the gene (whenever this is possible). It should be noted that specific arrays can detect splice variants by having probes detect specific exons or exon junctions (Castle, Garrett-Engele et al. 2003; Gardina, Clark et al. 2006; Bumgarner 2013).

Finally, microarrays measure, by design, relative concentration indirectly. The intensity measured in a probe, is proportional to the concentration of a sequence that can hybridize to this probe. However, experimental spike-in studies (Affymetrix 2001) showed that the probe intensity is nonlinearly proportional to the target concentration

(Chudin, Walker et al. 2002; Hekstra, Taussig et al. 2003; Skvortsov, Abdueva et al. 2007). The array will become saturated at high target concentrations, while at low concentrations there will be no binding. The intensities are linear within a very limited range of RNA concentration.

3.7 Microarray technology in the future

Technology that detects directly DNA or RNA sequences, such as NGS, will be much more preferable in the future. The massive decrease of sequencing cost has made NGS comparable in terms of cost with the microarrays (at the moment of writing, NGS is even cheaper for a few assays). Thus with similar costs, sequencing has several advantages relative to microarrays. Sequencing measures directly which nucleic acids are present in a sample and you only have to count the frequency of occurrence of a sequence is present in the sample to determine its abundance. Other advantages include the signal-to-noise ration which is limited by the number of reads for each sample and that counting is linearly related with the sample concentration. Sequencing is also less biased than microarrays in measuring which sequence is present in the sample. Unlike microarrays, sequencing is independent of prior design (knowledge) of which sequences might be present. It can also reliably measure the expression of homologous gene sequences and novel splice forms that cannot be reliably detected on microarrays.

As a result of sequencing decreasing cost and the aforementioned advantages, microarrays are gradually replaced by NGS for almost every assay. A search on the GEO public repository for arrays deposited within a 300day time period (between 21-June-2017 and 22-July-2017) (Table 8) reveals that microarrays are in decline, even though they still cover a big proportion of the data deposited on the repository. We did not use NGS on this thesis because there is much lower number of NGS muscle related data produced compared to microarray ones.

21/06/2017 to 22/07/2017	Microarrays	Next generation sequencing
Series	334	415
Samples	12474	19280
Median series | samples per day	8 | 228	13 | 270

Table 8 | GEO high-through put data submissions. Microarray and NGS data submitted in GEO during a 30 day period. Although microarrays are still used en-mass nowadays, the samples submitted are in decline as next generation sequencing is becoming cheaper.

Chapter 4 – References

(2015). "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." Science **348**(6235): 648-660.

Achanzar, W. E. and S. Ward (1997). "A nematode gene required for sperm vesicle fusion." J Cell Sci **110 (Pt 9)**: 1073-1081.

Adler, D., M. D., et al. (2016). "rgl: 3D Visualization Using OpenGL." R package version 0.95.1441.

Affymetrix. (2001). "Latin Square data for expression algorithm assessment." from https://www.thermofisher.com/fr/fr/home/life-science/microarray-analysis/microarray-data-analysis/microarray-analysis-sample-data/latin-square-data-expression-algorithm-assessment.html.

Affymetrix. (2002). "Statistical Algorithms Description Document." from http://www.affymetrix.com.

Affymetrix. (2004). "Expression Analysis Technical Manual." from www.affymetrix.com.

Affymetrix. (2006). "Affymetrix Power Tools." Retrieved 2017/05/15, from https://www.thermofisher.com/de/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html.

Affymetrix. (2009). "Affymetrix Data File Formats." from http://www.affymetrix.com.

Aken, B. L., S. Ayling, et al. (2016). "The Ensembl gene annotation system." Database (Oxford) **2016**.

Altman, N. S. (1992). "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." The American Statistician **46**(3): 175-185.

Amaratunga, D. and J. Cabrera (2001). "Analysis of Data From Viral DNA Microchips." Journal of the American Statistical Association **96**(456): 1161-1170.

Ampong, B. N., M. Imamura, et al. (2005). "Intracellular localization of dysferlin and its association with the dihydropyridine receptor." Acta Myol **24**(2): 134-144.

Anderson, L. V., K. Davison, et al. (1999). "Dysferlin is a plasma membrane protein and is expressed early in human development." Hum Mol Genet **8**(5): 855-861.

Anderson, L. V., R. M. Harrison, et al. (2000). "Secondary reduction in calpain 3 expression in patients with limb girdle muscular dystrophy type 2B and Miyoshi myopathy (primary dysferlinopathies)." Neuromuscul Disord **10**(8): 553-559.

Ankala, A., C. da Silva, et al. (2015). "A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield." Ann Neurol **77**(2): 206-214.

Aoki, K., Y. Ogata, et al. (2007). "Approaches for extracting practical information from gene co-expression networks in plant biology." Plant Cell Physiol **48**(3): 381-390.

Aoki, M., J. Liu, et al. (2001). "Genomic organization of the dysferlin gene and novel mutations in Miyoshi myopathy." Neurology **57**(2): 271-278.

Aoki, Y., Y. Okamura, et al. (2016). "ATTED-II in 2016: A Plant Coexpression Database Towards Lineage-Specific Coexpression." Plant Cell Physiol **57**(1): e5.

Aranda, B., H. Blankenburg, et al. (2011). "PSICQUIC and PSISCORE: accessing and scoring molecular interactions." Nat Methods **8**(7): 528-529.

Assadi, M., T. Schindler, et al. (2008). "Identification of proteins interacting with dysferlin using the tandem affinity purification method." Open Cell Dev. Biol. J **1**: 17-23.

Bakay, M., Z. Wang, et al. (2006). "Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration." Brain **129**(Pt 4): 996-1013.

Balci, B., G. Uyanik, et al. (2005). "An autosomal recessive limb girdle muscular dystrophy (LGMD2) with mild mental retardation is allelic to Walker-Warburg syndrome (WWS) caused by a mutation in the POMT1 gene." Neuromuscul Disord **15**(4): 271-275.

Baldi, P. and A. D. Long (2001). "A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes." Bioinformatics **17**(6): 509-519.

Ball, C. A. and A. Brazma (2006). "MGED standards: work in progress." OMICS **10**(2): 138-144.

Bansal, D. and K. P. Campbell (2004). "Dysferlin and the plasma membrane repair in muscular dystrophy." Trends Cell Biol **14**(4): 206-213.

Bansal, D., K. Miyake, et al. (2003). "Defective membrane repair in dysferlin-deficient muscular dystrophy." Nature **423**(6936): 168-172.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucleic Acids Res **35**(Database issue): D760-765.

Barrett, T., S. E. Wilhite, et al. (2013). "NCBI GEO: archive for functional genomics data sets--update." Nucleic Acids Res **41**(Database issue): D991-995.

Barthelemy, F., C. Blouin, et al. (2015). "Exon 32 Skipping of Dysferlin Rescues Membrane Repair in Patients' Cells." J Neuromuscul Dis **2**(3): 281-290.

Bashir, R., S. Britton, et al. (1998). "A gene related to Caenorhabditis elegans spermatogenesis factor fer-1 is mutated in limb-girdle muscular dystrophy type 2B." Nat Genet **20**(1): 37-42.

Bashir, R., T. Strachan, et al. (1994). "A gene for autosomal recessive limb-girdle muscular dystrophy maps to chromosome 2p." Hum Mol Genet **3**(3): 455-457.

Bejaoui, K., K. Hirabayashi, et al. (1995). "Linkage of Miyoshi myopathy (distal autosomal recessive muscular dystrophy) locus to chromosome 2p12-14." Neurology **45**(4): 768-772.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing " Journal of the Royal Statistical Society **57**: 11.

Bernard, C. C. and P. R. Carnegie (1975). "Experimental autoimmune encephalomyelitis in mice: immunologic response to mouse spinal cord and myelin basic proteins." J Immunol **114**(5): 1537-1540.

Beroud, C., G. Collod-Beroud, et al. (2000). "UMD (Universal mutation database): a generic software to build and analyze locus-specific databases." Hum Mutat **15**(1): 86-94.

Beroud, C., D. Hamroun, et al. (2005). "UMD (Universal Mutation Database): 2005 update." Hum Mutat **26**(3): 184-191.

Beurg, M., N. Michalski, et al. (2010). "Control of exocytosis by synaptotagmins and otoferlin in auditory hair cells." J Neurosci **30**(40): 13281-13290.

Bi, G. Q., J. M. Alderton, et al. (1995). "Calcium-regulated exocytosis is required for cell membrane resealing." J Cell Biol **131**(6 Pt 2): 1747-1758.

Bi, G. Q., R. L. Morris, et al. (1997). "Kinesin- and myosin-driven steps of vesicle recruitment for Ca2+-regulated exocytosis." J Cell Biol **138**(5): 999-1008.

Biancalana, V. and J. Laporte (2015). "Diagnostic use of Massively Parallel Sequencing in Neuromuscular Diseases: Towards an Integrated Diagnosis." J Neuromuscul Dis **2**(3): 193-203.

Biancheri, R., A. Falace, et al. (2007). "POMT2 gene mutation in limb-girdle muscular dystrophy with inflammatory changes." Biochem Biophys Res Commun **363**(4): 1033-1037.

Bisceglia, L., S. Zoccolella, et al. (2010). "A new locus on 3p23-p25 for an autosomal-dominant limb-girdle muscular dystrophy, LGMD1H." Eur J Hum Genet **18**(6): 636-641.

Bittner, R. E., L. V. Anderson, et al. (1999). "Dysferlin deletion in SJL mice (SJL-Dysf) defines a natural model for limb girdle muscular dystrophy 2B." Nat Genet **23**(2): 141-142.

Blake, D. J., A. Weir, et al. (2002). "Function and genetics of dystrophin and dystrophin-related proteins in muscle." Physiol Rev **82**(2): 291-329.

Blanchard, A. P., R. J. Kaiser, et al. (1996). "High-density oligonucleotide arrays." Biosensors and Bioelectronics **11**(6): 687-690.

Blandin, G., C. Beroud, et al. (2012). "UMD-DYSF, a novel locus specific database for the compilation and interactive analysis of mutations in the dysferlin gene." Hum Mutat **33**(3): E2317-2331.

Boedigheimer, M. J., R. D. Wolfinger, et al. (2008). "Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories." BMC Genomics **9**: 285.

Bogershausen, N., N. Shahrzad, et al. (2013). "Recessive TRAPPC11 mutations cause a disease spectrum of limb girdle muscular dystrophy and myopathy with movement disorder and intellectual disability." Am J Hum Genet **93**(1): 181-190.

Bolduc, V., G. Marlow, et al. (2010). "Recessive mutations in the putative calcium-activated chloride channel Anoctamin 5 cause proximal LGMD2L and distal MMD3 muscular dystrophies." Am J Hum Genet **86**(2): 213-221.

Bolstad, B. M., F. Collin, et al. (2005). Quality Assessment of Affymetrix GeneChip Data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry and S. Dudoit. New York, Springer.

Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Brayer, K. and J. L. Hammond, Jr. (1975). Evaluation of error detection polynomial performance on the AUTOVON channel. IEEE National Telecommunications Conference. New Orleans, LA, Institute of Electrical and Electronics Engineers. **1:** 8-21 to 28-25.

Brazma, A., P. Hingamp, et al. (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nat Genet **29**(4): 365-371.

Brazma, A., M. Kapushesky, et al. (2006). "Data storage and analysis in ArrayExpress." Methods Enzymol **411**: 370-386.

Brazma, A., H. Parkinson, et al. (2003). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res **31**(1): 68-71.

Brazma, A., A. Robinson, et al. (2000). "One-stop shop for microarray data." Nature **403**(6771): 699-700.

Brettschneider, J., F. Collin, et al. (2007). "Quality assessment for short oligonucleotide arrays." Technometrics.

Britton, S., T. Freeman, et al. (2000). "The third human FER-1-like protein is highly similar to dysferlin." Genomics **68**(3): 313-321.

Brockington, M., D. J. Blake, et al. (2001). "Mutations in the fukutin-related protein gene (FKRP) cause a form of congenital muscular dystrophy with secondary laminin alpha2 deficiency and abnormal glycosylation of alpha-dystroglycan." Am J Hum Genet **69**(6): 1198-1209.

Brown, P. O. and D. Botstein (1999). "Exploring the new world of the genome with DNA microarrays." Nat Genet **21**(1 Suppl): 33-37.

Bumgarner, R. (2013). "Overview of DNA microarrays: types, applications, and their future." Curr Protoc Mol Biol **Chapter 22**: Unit 22 21.

Bushby, K. M. (1999). "The limb-girdle muscular dystrophies-multiple genes, multiple mechanisms." Hum Mol Genet **8**(10): 1875-1882.

Bushby, K. M. D. (1999). "The Limb-Girdle Muscular Dystrophies—Multiple Genes, Multiple Mechanisms." Human Molecular Genetics **8**(10): 1875-1882.

Butte, A. (2002). "The use and analysis of microarray data." Nat Rev Drug Discov **1**(12): 951-960.

Cagliani, R., F. Fortunato, et al. (2003). "Molecular analysis of LGMD-2B and MM patients: identification of novel DYSF mutations and possible founder effect in the Italian population." Neuromuscular Disorders **13**(10): 788-795.

Cagliani, R., F. Magri, et al. (2005). "Mutation finding in patients with dysferlin deficiency and role of the dysferlin interacting proteins annexin A1 and A2 in muscular dystrophies." Hum Mutat **26**(3): 283.

Calderone, A., L. Castagnoli, et al. (2013). "mentha: a resource for browsing integrated protein-interaction networks." Nat Methods **10**(8): 690-691.

Campanaro, S., C. Romualdi, et al. (2002). "Gene expression profiling in dysferlinopathies using a dedicated muscle microarray." Hum Mol Genet **11**(26): 3283-3298.
Carlson, M. (2016). "hgfocus.db: Affymetrix Human Genome Focus Array annotation data (chip hgfocus)." R package version 3.2.3.
Carlson, M. (2016). "mouse4302.db: Affymetrix Mouse Genome 430 2.0 Array annotation data (chip mouse4302)." R package version 3.2.3.
Carss, K. J., E. Stevens, et al. (2013). "Mutations in GDP-mannose pyrophosphorylase B cause congenital and limb-girdle muscular dystrophies associated with hypoglycosylation of alpha-dystroglycan." Am J Hum Genet **93**(1): 29-41.
Castle, J., P. Garrett-Engele, et al. (2003). "Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing." Genome Biol **4**(10): R66.
Cetin, N., B. Balci-Hayta, et al. (2013). "A novel desmin mutation leading to autosomal recessive limb-girdle muscular dystrophy: distinct histopathological outcomes compared with desminopathies." J Med Genet **50**(7): 437-443.
Chapman, E. R. and R. Jahn (1994). "Calcium-dependent interaction of the cytoplasmic region of synaptotagmin with membranes. Autonomous function of a single C2-homologous domain." J Biol Chem **269**(8): 5735-5741.
Chardon, J. W., A. C. Smith, et al. (2015). "LIMS2 mutations are associated with a novel muscular dystrophy, severe cardiomyopathy and triangular tongues." Clin Genet **88**(6): 558-564.
Chase, T. H., G. A. Cox, et al. (2009). "Dysferlin deficiency and the development of cardiomyopathy in a mouse model of limb-girdle muscular dystrophy 2B." Am J Pathol **175**(6): 2299-2308.
Choe, S. E., M. Boutros, et al. (2005). "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset." Genome Biol **6**(2): R16.
Chudin, E., R. Walker, et al. (2002). "Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip(®) arrays." Genome Biology **3**(1): research0005.0001-research0005.0010.
Clark, N. R., R. Dannenfelser, et al. (2012). "Sets2Networks: network inference from repeated observations of sets." BMC Syst Biol **6**: 89.
Clark, N. R., K. S. Hu, et al. (2014). "The characteristic direction: a geometrical approach to identify differentially expressed genes." BMC Bioinformatics **15**: 79.
Clement, E. M., C. Godfrey, et al. (2008). "Mild POMGnT1 mutations underlie a novel limb-girdle muscular dystrophy variant." Arch Neurol **65**(1): 137-141.
Cohn, R. D. and K. P. Campbell (2000). "Molecular basis of muscular dystrophies." Muscle Nerve **23**(10): 1456-1471.
Cooper, S. T. and S. I. Head (2015). "Membrane Injury and Repair in the Muscular Dystrophies." Neuroscientist **21**(6): 653-668.
Cope, L. M., R. A. Irizarry, et al. (2004). "A benchmark for Affymetrix GeneChip expression measures." Bioinformatics **20**(3): 323-331.

Coral-Vazquez, R., R. D. Cohn, et al. (1999). "Disruption of the sarcoglycan-sarcospan complex in vascular smooth muscle: a novel mechanism for cardiomyopathy and muscular dystrophy." Cell **98**(4): 465-474.

Cortes, C. and V. Vapnik (1995). "Support-vector networks." Machine Learning **20**(3): 273-297.

D'Haeseleer, P. (2005). "How does gene expression clustering work?" Nat Biotech **23**(12): 1499-1501.

Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucleic Acids Res **33**(20): e175.

Dalma-Weiszhausz, D. D., J. Warrington, et al. (2006). "The affymetrix GeneChip platform: an overview." Methods Enzymol **410**: 3-28.

Davis, D. B., A. J. Delmonte, et al. (2000). "Myoferlin, a candidate gene and potential modifier of muscular dystrophy." Hum Mol Genet **9**(2): 217-226.

Davis, D. B., K. R. Doherty, et al. (2002). "Calcium-sensitive phospholipid binding properties of normal and mutant ferlin C2 domains." J Biol Chem **277**(25): 22883-22888.

Davletov, B. A. and T. C. Sudhof (1993). "A single C2 domain from synaptotagmin I is sufficient for high affinity Ca2+/phospholipid binding." J Biol Chem **268**(35): 26386-26390.

de la Fuente, A. (2010). "From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases." Trends Genet **26**(7): 326-333.

De Luna, N., A. Freixas, et al. (2007). "Dysferlin expression in monocytes: a source of mRNA for mutation analysis." Neuromuscul Disord **17**(1): 69-76.

de Luna, N., E. Gallardo, et al. (2006). "Absence of dysferlin alters myogenin expression and delays human muscle differentiation "in vitro"." J Biol Chem **281**(25): 17092-17098.

De Smet, R. and K. Marchal (2010). "Advantages and limitations of current network inference methods." Nat Rev Microbiol **8**(10): 717-729.

de Souto, M. C., I. G. Costa, et al. (2008). "Clustering cancer gene expression data: a comparative study." BMC Bioinformatics **9**(1): 497.

Demonbreun, A. R., M. V. Allen, et al. (2016). "Enhanced Muscular Dystrophy from Loss of Dysferlin Is Accompanied by Impaired Annexin A6 Translocation after Sarcolemmal Disruption." Am J Pathol **186**(6): 1610-1622.

Demonbreun, A. R., J. P. Fahrenbach, et al. (2011). "Impaired muscle growth and response to insulin-like growth factor 1 in dysferlin-mediated muscular dystrophy." Hum Mol Genet **20**(4): 779-789.

DeRisi, J., L. Penland, et al. (1996). "Use of a cDNA microarray to analyse gene expression patterns in human cancer." Nat Genet **14**(4): 457-460.

Derrien, T., R. Johnson, et al. (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." Genome Res **22**(9): 1775-1789.

Deutsch, E. W., C. A. Ball, et al. (2008). "Minimum information specification for in situ hybridization and immunohistochemistry experiments (MISFISHIE)." Nat Biotechnol **26**(3): 305-312.
Diers, A., M. Carl, et al. (2007). "Painful enlargement of the calf muscles in limb girdle muscular dystrophy type 2B (LGMD2B) with a novel compound heterozygous mutation in DYSF." Neuromuscul Disord **17**(2): 157-162.
Doherty, K. R., A. Cave, et al. (2005). "Normal myoblast fusion requires myoferlin." Development **132**(24): 5565-5575.
Eastlake, D. (2001). "Secure Hash Algorithm 1 (SHA1)."
Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.
Emmert-Streib, F. and M. Dehmer (2008). Analysis of Microarray Data A Network-Based Approach, Wiley-VCH.
Engelman, J. A., X. Zhang, et al. (1998). "Molecular genetics of the caveolin gene family: implications for human cancers, diabetes, Alzheimer disease, and muscular dystrophy." Am J Hum Genet **63**(6): 1578-1587.
Evans, W. E. and M. V. Relling (2004). "Moving towards individualized medicine with pharmacogenomics." Nature **429**(6990): 464-468.
Fabregat, A., K. Sidiropoulos, et al. (2016). "The Reactome pathway Knowledgebase." Nucleic Acids Res **44**(D1): D481-487.
Fanin, M., A. C. Nascimbeni, et al. (2005). "The frequency of limb girdle muscular dystrophy 2A in northeastern Italy." Neuromuscul Disord **15**(3): 218-224.
Ferguson, J. A., F. J. Steemers, et al. (2000). "High-density fiber-optic DNA random microsphere array." Anal Chem **72**(22): 5618-5624.
Fernandez, I., D. Arac, et al. (2001). "Three-dimensional structure of the synaptotagmin 1 C2B-domain: synaptotagmin 1 as a phospholipid binding machine." Neuron **32**(6): 1057-1069.
Field, D., G. Garrity, et al. (2008). "The minimum information about a genome sequence (MIGS) specification." Nat Biotechnol **26**(5): 541-547.
Fisher, R. A. (1915). "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population." Biometrika **10**(4): 507-521.
Florez-Vargas, O., A. Brass, et al. (2016). "Bias in the reporting of sex and age in biomedical research on mouse models." Elife **5**.
Fodor, S. P., J. L. Read, et al. (1991). "Light-directed, spatially addressable parallel chemical synthesis." Science **251**(4995): 767-773.
Forgy, E. (1965). "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications." Biometrics **21**(3): 761-777.
Frosk, P., T. Weiler, et al. (2002). "Limb-girdle muscular dystrophy type 2H associated with mutation in TRIM32, a putative E3-ubiquitin-ligase gene." Am J Hum Genet **70**(3): 663-672.
Fuson, K., A. Rice, et al. (2014). "Alternate splicing of dysferlin C2A confers Ca(2)(+)-dependent and Ca(2)(+)-independent binding for membrane repair." Structure **22**(1): 104-115.

Futamura, N., Y. Nishida, et al. (2014). "EMMPRIN co-expressed with matrix metalloproteinases predicts poor prognosis in patients with osteosarcoma." Tumour Biol **35**(6): 5159-5165.
Galbiati, F., B. Razani, et al. (2001). "Caveolae and caveolin-3 in muscular dystrophy." Trends Mol Med **7**(10): 435-441.
Gardina, P. J., T. A. Clark, et al. (2006). "Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array." BMC Genomics **7**: 325.
Gene Ontology Consortium (2015). "Gene Ontology Consortium: going forward." Nucleic Acids Res **43**(Database issue): D1049-1056.
Gentleman, R. C., V. J. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.
Gerhold, D. L., R. V. Jensen, et al. (2002). "Better therapeutics through microarrays." Nat Genet **32 Suppl**: 547-551.
Gerke, V., C. E. Creutz, et al. (2005). "Annexins: linking Ca2+ signalling to membrane dynamics." Nat Rev Mol Cell Biol **6**(6): 449-461.
Godfrey, C., D. Escolar, et al. (2006). "Fukutin gene mutations in steroid-responsive limb girdle muscular dystrophy." Ann Neurol **60**(5): 603-610.
Gower, J. C. (1966). "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." Biometrika **53**(3/4): 325-338.
Greenberg, S. A., M. Salajegheh, et al. (2012). "Etiology of limb girdle muscular dystrophy 1D/1E determined by laser capture microdissection proteomics." Ann Neurol **71**(1): 141-145.
Greene, C. S., A. Krishnan, et al. (2015). "Understanding multicellular function and disease with human tissue-specific networks." Nat Genet **47**(6): 569-576.
Grounds, M. D. and J. K. McGeachie (1989). "A comparison of muscle precursor replication in crush-injured skeletal muscle of Swiss and BALBc mice." Cell Tissue Res **255**(2): 385-391.
Grunstein, M. and D. S. Hogness (1975). "Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene." Proc Natl Acad Sci U S A **72**(10): 3961-3965.
Guglieri, M., V. Straub, et al. (2008). "Limb-girdle muscular dystrophies." Curr Opin Neurol **21**(5): 576-584.
Gundesli, H., B. Talim, et al. (2010). "Mutation in exon 1f of PLEC, leading to disruption of plectin isoform 1f, causes autosomal-recessive limb-girdle muscular dystrophy." Am J Hum Genet **87**(6): 834-841.
Hackman, P., A. Vihola, et al. (2002). "Tibial muscular dystrophy is a titinopathy caused by mutations in TTN, the gene encoding the giant skeletal-muscle protein titin." Am J Hum Genet **71**(3): 492-500.
Han, R. and K. P. Campbell (2007). "Dysferlin and muscle membrane repair." Curr Opin Cell Biol **19**(4): 409-416.
Hand, D. J. and K. Yu (2001). "Idiot's Bayes: Not So Stupid after All?" International Statistical Review / Revue Internationale de Statistique **69**(3): 385-398.

Hara, Y., B. Balci-Hayta, et al. (2011). "A dystroglycan mutation associated with limb-girdle muscular dystrophy." N Engl J Med **364**(10): 939-946.
Harms, M. B., R. B. Sommerville, et al. (2012). "Exome sequencing reveals DNAJB6 mutations in dominantly-inherited myopathy." Ann Neurol **71**(3): 407-416.
Hauser, M. A., S. K. Horrigan, et al. (2000). "Myotilin is mutated in limb girdle muscular dystrophy 1A." Hum Mol Genet **9**(14): 2141-2147.
He, L., M. Vanlandewijck, et al. (2016). "Analysis of the brain mural cell transcriptome." Sci Rep **6**: 35108.
Hedberg, C., A. Melberg, et al. (2012). "Autosomal dominant myofibrillar myopathy with arrhythmogenic right ventricular cardiomyopathy 7 is caused by a DES mutation." Eur J Hum Genet **20**(9): 984-985.
Hekstra, D., A. R. Taussig, et al. (2003). "Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays." Nucleic Acids Research **31**(7): 1962-1968.
Ho, M., C. M. Post, et al. (2004). "Disruption of muscle membrane and phenotype divergence in two novel mouse models of dysferlin deficiency." Hum Mol Genet **13**(18): 1999-2010.
Hochreiter, S., D. A. Clevert, et al. (2006). "A new summarization method for Affymetrix probe level data." Bioinformatics **22**(8): 943-949.
Hoffman, E. P., R. H. Brown, Jr., et al. (1987). "Dystrophin: the protein product of the Duchenne muscular dystrophy locus." Cell **51**(6): 919-928.
Huang, Y., S. H. Laval, et al. (2007). "AHNAK, a novel component of the dysferlin protein complex, redistributes to the cytoplasm with dysferlin during skeletal muscle regeneration." FASEB J **21**(3): 732-742.
Huang, Y., P. Verheesen, et al. (2005). "Protein studies in dysferlinopathy patients using llama-derived antibody fragments selected by phage display." Eur J Hum Genet **13**(6): 721-730.
Hubbell, E. (2005). "Affymetrix technical notes: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation." from http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf.
Hubbell, E. (2005). "Affymetrix white paper: Gene Signal Estimates from Exon Arrays." from http://www.affymetrix.com/support/technical/whitepapers/exon_gene_signal_estimate_whitepaper.pdf.
Hubbell, E., W. M. Liu, et al. (2002). "Robust estimators for expression analysis." Bioinformatics **18**(12): 1585-1592.
Huber, W., V. J. Carey, et al. (2015). "Orchestrating high-throughput genomic analysis with Bioconductor." Nat Methods **12**(2): 115-121.
Hudson, N. J., A. Reverter, et al. (2009). "A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation." PLoS Comput Biol **5**(5): e1000382.
Hughes, T. R., M. Mao, et al. (2001). "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer." Nat Biotechnol **19**(4): 342-347.

Humphrey, G. W., E. Mekhedov, et al. (2012). "GREG cells, a dysferlin-deficient myogenic mouse cell line." Exp Cell Res **318**(2): 127-135.
Illa, I., N. De Luna, et al. (2007). "Symptomatic dysferlin gene mutation carriers: characterization of two cases." Neurology **68**(16): 1284-1289.
Illa, I., C. Serrano-Munuera, et al. (2001). "Distal anterior compartment myopathy: a dysferlin mutation causing a new muscular dystrophy phenotype." Ann Neurol **49**(1): 130-134.
Inoue, M., Y. Wakayama, et al. (2006). "Expression of myoferlin in skeletal muscles of patients with dysferlinopathy." Tohoku J Exp Med **209**(2): 109-116.
Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Res **31**(4): e15.
Irizarry, R. A., B. Hobbs, et al. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.
Irizarry, R. A., D. Warren, et al. (2005). "Multiple-laboratory comparison of microarray platforms." Nat Methods **2**(5): 345-350.
Irizarry, R. A., Z. Wu, et al. (2006). "Comparison of Affymetrix GeneChip expression measures." Bioinformatics **22**(7): 789-794.
Jen, C. H., I. W. Manfield, et al. (2006). "The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis." Plant J **46**(2): 336-348.
Jiang, N., L. J. Leach, et al. (2008). "Methods for evaluating gene expression from Affymetrix microarray datasets." BMC Bioinformatics **9**(1): 284.
Johnson, C. P. and E. R. Chapman (2010). "Otoferlin is a calcium sensor that directly regulates SNARE-mediated membrane fusion." J Cell Biol **191**(1): 187-197.
Johnson, W. E., C. Li, et al. (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." Biostatistics **8**(1): 118-127.
Jupiter, D., H. Chen, et al. (2009). "STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data." BMC Bioinformatics **10**: 332.
Kanehisa, M., M. Furumichi, et al. (2017). "KEGG: new perspectives on genomes, pathways, diseases and drugs." Nucleic Acids Res **45**(D1): D353-D361.
Kell, D. B. and S. G. Oliver (2004). "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era." Bioessays **26**(1): 99-105.
Kerr, J. P., A. P. Ziman, et al. (2013). "Dysferlin stabilizes stress-induced Ca2+ signaling in the transverse tubule membrane." Proc Natl Acad Sci U S A **110**(51): 20831-20836.
Kim, M. S., S. M. Pinto, et al. (2014). "A draft map of the human proteome." Nature **509**(7502): 575-581.
Kinsella, R. J., A. Kahari, et al. (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space." Database (Oxford) **2011**: bar030.
Klinge, L., A. Aboumousa, et al. (2010). "New aspects on patients affected by dysferlin deficient muscular dystrophy." J Neurol Neurosurg Psychiatry **81**(9): 946-953.

Klinge, L., J. Harris, et al. (2010). "Dysferlin associates with the developing T-tubule system in rodent and human skeletal muscle." Muscle Nerve **41**(2): 166-173.
Klinge, L., S. Laval, et al. (2007). "From T-tubule to sarcolemma: damage-induced dysferlin translocation in early myogenesis." FASEB J **21**(8): 1768-1776.
Kohonen, T. (1982). "Self-organized formation of topologically correct feature maps." Biological Cybernetics **43**(1): 59-69.
Kolesnikov, N., E. Hastings, et al. (2015). "ArrayExpress update--simplifying data submissions." Nucleic Acids Res **43**(Database issue): D1113-1116.
Komuro, A., Y. Masuda, et al. (2004). "The AHNAKs are a class of giant propeller-like proteins that associate with calcium channel proteins of cardiomyocytes and other cells." Proc Natl Acad Sci U S A **101**(12): 4053-4058.
Kostka, D. and R. Spang (2004). "Finding disease specific alterations in the co-expression of genes." Bioinformatics **20 Suppl 1**: i194-199.
Krahn, M., C. Beroud, et al. (2009). "Analysis of the DYSF mutational spectrum in a large cohort of patients." Hum Mutat **30**(2): E345-375.
Krajacic, P., J. Hermanowski, et al. (2009). "C. elegans dysferlin homolog fer-1 is expressed in muscle, and fer-1 mutations initiate altered gene expression of muscle enriched genes." Physiol Genomics **40**(1): 8-14.
Kristiansson, E., A. Sjogren, et al. (2006). "Quality optimised analysis of general paired microarray experiments." Stat Appl Genet Mol Biol **5**: Article10.
Kuleshov, M. V., M. R. Jones, et al. (2016). "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update." Nucleic Acids Res **44**(W1): W90-97.
Kundaje, A., W. Meuleman, et al. (2015). "Integrative analysis of 111 reference human epigenomes." Nature **518**(7539): 317-330.
Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." BMC Bioinformatics **9**: 559.
Larance, M. and A. I. Lamond (2015). "Multidimensional proteomics for cell biology." Nat Rev Mol Cell Biol **16**(5): 269-280.
Laval, S. H. and K. M. Bushby (2004). "Limb-girdle muscular dystrophies--from genetics to molecular pathology." Neuropathol Appl Neurobiol **30**(2): 91-105.
Lee, Y. S., A. Lehar, et al. (2015). "Muscle hypertrophy induced by myostatin inhibition accelerates degeneration in dysferlinopathy." Hum Mol Genet **24**(20): 5711-5719.
Leek, J. T. (2014). "svaseq: removing batch effects and other unwanted noise from sequencing data." Nucleic Acids Res **42**(21).
Leek, J. T., W. E. Johnson, et al. (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." Bioinformatics **28**(6): 882-883.
Leek, J. T., R. B. Scharpf, et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." Nat Rev Genet **11**(10): 733-739.
Lek, A., M. Lek, et al. (2010). "Phylogenetic analysis of ferlin genes reveals ancient eukaryotic origins." BMC Evol Biol **10**: 231.

Lek, M. and D. MacArthur (2014). "The Challenge of Next Generation Sequencing in the Context of Neuromuscular Diseases." J Neuromuscul Dis **1**(2): 135-149.
Lemmers, R. J., M. Wohlgemuth, et al. (2007). "Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy." Am J Hum Genet **81**(5): 884-894.
Lennon, N. J., A. Kho, et al. (2003). "Dysferlin interacts with annexins A1 and A2 and mediates sarcolemmal wound-healing." J Biol Chem **278**(50): 50466-50473.
Letunic, I., T. Doerks, et al. (2015). "SMART: recent updates, new developments and status in 2015." Nucleic Acids Res **43**(Database issue): D257-260.
Li, C. and W. H. Wong (2001). "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection." Proc Natl Acad Sci U S A **98**(1): 31-36.
Li, K. C. (2002). "Genome-wide coexpression dynamics: theory and application." Proc Natl Acad Sci U S A **99**(26): 16875-16880.
Lim, L. E., F. Duclos, et al. (1995). "Beta-sarcoglycan: characterization and role in limb-girdle muscular dystrophy linked to 4q12." Nat Genet **11**(3): 257-265.
Lim, W. K., K. Wang, et al. (2007). "Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks." Bioinformatics **23**(13): i282-288.
Lipshutz, R. J., S. P. Fodor, et al. (1999). "High density synthetic oligonucleotide arrays." Nat Genet **21**(1 Suppl): 20-24.
Liu, J., M. Aoki, et al. (1998). "Dysferlin, a novel skeletal muscle gene, is mutated in Miyoshi myopathy and limb girdle muscular dystrophy." Nat Genet **20**(1): 31-36.
Llanga, T., N. Nagy, et al. (2017). "Structure-Based Designed Nano-Dysferlin Significantly Improves Dysferlinopathy in BLA/J Mice." Mol Ther.
Lloyd, S. (1982). "Least squares quantization in PCM." IEEE Transactions on Information Theory **28**(2): 129-137.
Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-1680.
Lonnstedt, I. and S. T. P. (2002). "Replicated microarray data." Statistica Sinica **12**(1): 31–46.
Lostal, W., M. Bartoli, et al. (2010). "Efficient recovery of dysferlin deficiency by dual adeno-associated vector-mediated gene transfer." Hum Mol Genet **19**(10): 1897-1907.
Lovering, R. M., N. C. Porter, et al. (2005). "The muscular dystrophies: from genes to therapies." Phys Ther **85**(12): 1372-1388.
Lu, Z. and D. Shen "Computation of Correlation Coefficient and Its Confidence Interval in SAS."
Ma, R. L., L. Y. Shen, et al. (2014). "Coexpression of ANXA2, SOD2 and HOXA13 predicts poor prognosis of esophageal squamous cell carcinoma." Oncol Rep **31**(5): 2157-2164.
Ma, S., Q. Gong, et al. (2007). "An Arabidopsis gene network based on the graphical Gaussian model." Genome Res **17**(11): 1614-1625.

Maglott, D., J. Ostell, et al. (2011). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res **39**(Database issue): D52-57.
Mahjneh, I., K. Bushby, et al. (1996). "Limb-girdle muscular dystrophy: a follow-up study of 79 patients." Acta Neurol Scand **94**(3): 177-189.
Mahjneh, I., G. Marconi, et al. (2001). "Dysferlinopathy (LGMD2B): a 23-year follow-up study of 10 patients homozygous for the same frameshifting dysferlin mutations." Neuromuscul Disord **11**(1): 20-26.
Mahjneh, I., G. Vannelli, et al. (1992). "A large inbred Palestinian family with two forms of muscular dystrophy." Neuromuscul Disord **2**(4): 277-283.
Marbach, D., J. C. Costello, et al. (2012). "Wisdom of crowds for robust gene network inference." Nat Methods **9**(8): 796-804.
Margolin, A. A., I. Nemenman, et al. (2006). "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." BMC Bioinformatics **7 Suppl 1**: S7.
Matsuda, C., Y. K. Hayashi, et al. (2001). "The sarcolemmal proteins dysferlin and caveolin-3 interact in skeletal muscle." Hum Mol Genet **10**(17): 1761-1766.
Matsuda, C., K. Kameyama, et al. (2005). "Dysferlin interacts with affixin (beta-parvin) at the sarcolemma." J Neuropathol Exp Neurol **64**(4): 334-340.
McCall, M. N., P. N. Murakami, et al. (2011). "Assessing affymetrix GeneChip microarray quality." BMC Bioinformatics **12**: 137.
McGeachie, J. K. and M. D. Grounds (1995). "Retarded myogenic cell replication in regenerating skeletal muscles of old mice: an autoradiographic study in young and old BALBc and SJL/J mice." Cell Tissue Res **280**(2): 277-282.
McNally, E. M., E. de Sa Moreira, et al. (1998). "Caveolin-3 in muscular dystrophy." Hum Mol Genet **7**(5): 871-877.
McNeil, A. K., U. Rescher, et al. (2006). "Requirement for annexin A1 in plasma membrane repair." J Biol Chem **281**(46): 35202-35207.
McNeil, P. L., K. Miyake, et al. (2003). "The endomembrane requirement for cell surface repair." Proc Natl Acad Sci U S A **100**(8): 4592-4597.
McNeil, P. L. and R. A. Steinhardt (1997). "Loss, restoration, and maintenance of plasma membrane integrity." J Cell Biol **137**(1): 1-4.
Meldolesi, J. (2003). "Surface wound healing: a new, general function of eukaryotic cells." J Cell Mol Med **7**(3): 197-203.
Melia, M. J., A. Kubota, et al. (2013). "Limb-girdle muscular dystrophy 1F is caused by a microdeletion in the transportin 3 gene." Brain **136**(Pt 5): 1508-1517.
Mercuri, E. and F. Muntoni (2012). "The ever-expanding spectrum of congenital muscular dystrophies." Ann Neurol **72**(1): 9-17.
Michael, K. L., L. C. Taylor, et al. (1998). "Randomly ordered addressable high-density optical sensor arrays." Anal Chem **70**(7): 1242-1248.
Michalopoulos, I., G. A. Pavlopoulos, et al. (2012). "Human gene correlation analysis (HGCA): A tool for the identification of transcriptionally coexpressed genes." BMC Res Notes **5**(1): 265.
Middel, V., L. Zhou, et al. (2016). "Dysferlin-mediated phosphatidylserine sorting engages macrophages in sarcolemma repair." Nat Commun **7**: 12875.

Miller, C. J. (2017). "simpleaffy: Very simple high level analysis of Affymetrix data."
Miller, M. B. and Y. W. Tang (2009). "Basic concepts of microarrays and potential applications in clinical microbiology." Clin Microbiol Rev **22**(4): 611-633.
Minetti, C., F. Sotgia, et al. (1998). "Mutations in the caveolin-3 gene cause autosomal dominant limb-girdle muscular dystrophy." Nat Genet **18**(4): 365-368.
Mitchell, C. A., J. K. McGeachie, et al. (1992). "Cellular differences in the regeneration of murine skeletal muscle: a quantitative histological study in SJL/J and BALB/c mice." Cell Tissue Res **269**(1): 159-166.
Miyoshi, K., H. Kawai, et al. (1986). "Autosomal recessive distal muscular dystrophy as a new type of progressive muscular dystrophy. Seventeen cases in eight families including an autopsied case." Brain **109 (Pt 1)**: 31-54.
Miyoshi, K., K. Saijo, et al. (1967). "Four cases of distal myopathy in two families." Jpn. J. Hum. Genet. **12**: 113.
Moore, S. A., F. Saito, et al. (2002). "Deletion of brain dystroglycan recapitulates aspects of congenital muscular dystrophy." Nature **418**(6896): 422-425.
Moreira, E. S., T. J. Wiltshire, et al. (2000). "Limb-girdle muscular dystrophy type 2G is caused by mutations in the gene encoding the sarcomeric protein telethonin." Nat Genet **24**(2): 163-166.
Muchir, A., G. Bonne, et al. (2000). "Identification of mutations in the gene encoding lamins A/C in autosomal dominant limb girdle muscular dystrophy with atrioventricular conduction disturbances (LGMD1B)." Hum Mol Genet **9**(9): 1453-1459.
Muntoni, F., S. Torelli, et al. (2011). "Muscular dystrophies due to glycosylation defects: diagnosis and therapeutic strategies." Curr Opin Neurol **24**(5): 437-442.
Naef, F., D. A. Lim, et al. (2001). "From features to expression: High density oligonucleotide array analysis revisited." Tech Report **1**: 1-9.
Naef, F., D. A. Lim, et al. (2002). "DNA hybridization to mismatched templates: a chip study." Phys Rev E Stat Nonlin Soft Matter Phys **65**(4 Pt 1): 040902.
Naef, F. and M. O. Magnasco (2003). "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays." Physical Review E **68**(1): 011906.
Narayanaswami, P., G. Carter, et al. (2015). "Evidence-based guideline summary: Diagnosis and treatment of limb-girdle and distal dystrophies: Report of the Guideline Development Subcommittee of the American Academy of Neurology and the Practice Issues Review Panel of the American Association of Neuromuscular & Electrodiagnostic Medicine." Neurology **84**(16): 1720-1721.
Newton, A. C. (1995). "Protein kinase C. Seeing two domains." Curr Biol **5**(9): 973-976.
Nguyen, K., G. Bassez, et al. (2005). "Dysferlin mutations in LGMD2B, Miyoshi myopathy, and atypical dysferlinopathies." Hum Mutat **26**(2): 165.
Nguyen, K., G. Bassez, et al. (2007). "Phenotypic study in 40 patients with dysferlin gene mutations: high frequency of atypical phenotypes." Arch Neurol **64**(8): 1176-1182.
Nigro, V., S. Aurino, et al. (2011). "Limb girdle muscular dystrophies: update on genetic diagnosis and therapeutic approaches." Curr Opin Neurol **24**(5): 429-436.

Nigro, V., E. de Sa Moreira, et al. (1996). "Autosomal recessive limb-girdle muscular dystrophy, LGMD2F, is caused by a mutation in the delta-sarcoglycan gene." Nat Genet **14**(2): 195-198.
Nigro, V. and M. Savarese (2014). "Genetic basis of limb-girdle muscular dystrophies: the 2014 update." Acta Myol **33**(1): 1-12.
Noguchi, S., E. M. McNally, et al. (1995). "Mutations in the dystrophin-associated protein gamma-sarcoglycan in chromosome 13 muscular dystrophy." Science **270**(5237): 819-822.
Nonaka, I., N. Sunohara, et al. (1981). "Familial distal myopathy with rimmed vacuole and lamellar (myeloid) body formation." J Neurol Sci **51**(1): 141-155.
Nygaard, V., E. A. Rodland, et al. (2016). "Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses." Biostatistics **17**(1): 29-39.
Okamura, Y., Y. Aoki, et al. (2015). "COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems." Nucleic Acids Res **43**(Database issue): D82-86.
Orchard, S., M. Ammari, et al. (2014). "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases." Nucleic Acids Research **42**(D1): D358-D363.
Ostlund, C. and H. J. Worman (2003). "Nuclear envelope proteins and neuromuscular diseases." Muscle Nerve **27**(4): 393-406.
Pakula, A., J. Schneider, et al. (2013). "Altered expression of cyclin A 1 in muscle of patients with facioscapulohumeral muscle dystrophy (FSHD-1)." PLoS One **8**(9): e73573.
Pallanck, L. (2003). "A tale of two C2 domains." Trends Neurosci **26**(1): 2-4.
Parker, C. E. and C. H. Borchers (2014). "Mass spectrometry based biomarker discovery, verification, and validation--quality assurance and control of protein biomarker assays." Mol Oncol **8**(4): 840-858.
Parman, C., C. Halling, et al. (2017). "affyQCReport: QC Report Generation for affyBatch objects." R package version 1.54.0.
Passamano, L., A. Taglia, et al. (2012). "Improvement of survival in Duchenne Muscular Dystrophy: retrospective analysis of 835 patients." Acta Myol **31**(2): 121-125.
Pearson, K. (1901). "LIII. On lines and planes of closest fit to systems of points in space." Philosophical Magazine **2**(11): 559-572.
Pearson, K. (1920). "Notes on the History of Correlation." Biometrika **13**: 25-45.
Pease, A. C., D. Solas, et al. (1994). "Light-generated oligonucleotide arrays for rapid DNA sequence analysis." Proc Natl Acad Sci U S A **91**(11): 5022-5026.
Petrof, B. J., J. B. Shrager, et al. (1993). "Dystrophin protects the sarcolemma from stresses developed during muscle contraction." Proc Natl Acad Sci U S A **90**(8): 3710-3714.
Petryszak, R., M. Keays, et al. (2016). "Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants." Nucleic Acids Res **44**(D1): D746-752.

Pettersson, E., J. Lundeberg, et al. (2009). "Generations of sequencing technologies." Genomics **93**(2): 105-111.
Philippi, S., A. Bigot, et al. (2012). "Dysferlin-deficient immortalized human myoblasts and myotubes as a useful tool to study dysferlinopathy." PLoS Curr **4**: RRN1298.
Piccolo, S. R., Y. Sun, et al. (2012). "A single-sample microarray normalization method to facilitate personalized-medicine workflows." Genomics **100**(6): 337-344.
Piccolo, S. R., M. R. Withers, et al. (2013). "Multiplatform single-sample estimates of transcriptional activation." Proc Natl Acad Sci U S A **110**(44): 17778-17783.
Piro, R. M., U. Ala, et al. (2011). "An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction." Eur J Hum Genet **19**(11): 1173-1180.
Potter, R. A., D. A. Griffin, et al. (2017). "Systemic Delivery of Dysferlin Overlap Vectors Provides Long-Term Functional Improvement for Dysferlinopathy." Hum Gene Ther.
Preisler, N., Z. Lukacs, et al. (2013). "Late-onset Pompe disease is prevalent in unclassified limb-girdle muscular dystrophies." Mol Genet Metab **110**(3): 287-289.
PyMOL The PyMOL Molecular Graphics System, Schrödinger, LLC.
Raynal, P. and H. B. Pollard (1994). "Annexins: the problem of assessing the biological role for a gene family of multifunctional calcium- and phospholipid-binding proteins." Biochim Biophys Acta **1197**(1): 63-93.
Rayner, T. F., P. Rocca-Serra, et al. (2006). "A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB." BMC Bioinformatics **7**: 489.
Reddy, A., E. V. Caler, et al. (2001). "Plasma membrane repair is mediated by Ca(2+)-regulated exocytosis of lysosomes." Cell **106**(2): 157-169.
Revelle, W. (2017). psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston, Illinois, USA.
Richard, I., O. Broux, et al. (1995). "Mutations in the proteolytic enzyme calpain 3 cause limb-girdle muscular dystrophy type 2A." Cell **81**(1): 27-40.
Ritchie, M. E., B. Phipson, et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Res **43**(7): e47.
Rizo, J. and T. C. Sudhof (1998). "C2-domains, structure and function of a universal Ca2+-binding domain." J Biol Chem **273**(26): 15879-15882.
Roberds, S. L., F. Leturcq, et al. (1994). "Missense mutations in the adhalin gene linked to autosomal recessive muscular dystrophy." Cell **78**(4): 625-633.
Rosenberg, N. L., S. P. Ringel, et al. (1987). "Experimental autoimmune myositis in SJL/J mice." Clin Exp Immunol **68**(1): 117-129.
Roux, I., S. Safieddine, et al. (2006). "Otoferlin, defective in a human deafness form, is essential for exocytosis at the auditory ribbon synapse." Cell **127**(2): 277-289.
Safran, M., I. Dalah, et al. (2010). "GeneCards Version 3: the human gene integrator." Database (Oxford) **2010**: baq020.
Sandberg, R. and O. Larsson (2007). "Improved precision and accuracy for microarrays using updated probe set definitions." BMC Bioinformatics **8**: 48.

Sarparanta, J., P. H. Jonson, et al. (2012). "Mutations affecting the cytoplasmic functions of the co-chaperone DNAJB6 cause limb-girdle muscular dystrophy." Nat Genet **44**(4): 450-455, S451-452.
Sartor, M. A., C. R. Tomlinson, et al. (2006). "Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments." BMC Bioinformatics **7**: 538.
Schena, M. (1996). "Genome analysis with gene expression microarrays." Bioessays **18**(5): 427-431.
Schena, M. (2002). Microarray Analysis. Hoboken, New Jersey, Wiley-Liss.
Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.
Schindler, R. F., C. Scotton, et al. (2016). "POPDC1(S201F) causes muscular dystrophy and arrhythmia by affecting protein trafficking." J Clin Invest **126**(1): 239-253.
Schuster, E. F., E. Blanc, et al. (2007). "Estimation and correction of non-specific binding in a large-scale spike-in experiment." Genome Biol **8**(6): R126.
Seror, P., M. Krahn, et al. (2008). "Complete fatty degeneration of lumbar erector spinae muscles caused by a primary dysferlinopathy." Muscle Nerve **37**(3): 410-414.
Sharma, A., C. Yu, et al. (2010). "A new role for the muscle repair protein dysferlin in endothelial cell adhesion and angiogenesis." Arterioscler Thromb Vasc Biol **30**(11): 2196-2204.
Simon, M. J., C. Murchison, et al. (2017). "A transcriptome-based assessment of the astrocytic dystrophin-associated complex in the developing human brain." J Neurosci Res.
Skvortsov, D., D. Abdueva, et al. (2007). "Explaining differences in saturation levels for Affymetrix GeneChip® arrays." Nucleic Acids Research **35**(12): 4154-4163.
Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Stat Appl Genet Mol Biol **3**: Article3.
Sondergaard, P. C., D. A. Griffin, et al. (2015). "AAV.Dysferlin Overlap Vectors Restore Function in Dysferlinopathy Animal Models." Ann Clin Transl Neurol **2**(3): 256-270.
Spearman, C. (1904). "The Proof and Measurement of Association between Two Things." The American Journal of Psychology **15**(1): 72-101.
Spellman, P. T., M. Miller, et al. (2002). "Design and implementation of microarray gene expression markup language (MAGE-ML)." Genome Biol **3**(9): RESEARCH0046.
Steemers, F. J., J. A. Ferguson, et al. (2000). "Screening unlabeled DNA targets with randomly ordered fiber-optic gene arrays." Nat Biotechnol **18**(1): 91-94.
Steuer, R., J. Kurths, et al. (2002). "The mutual information: detecting and evaluating dependencies between variables." Bioinformatics **18 Suppl 2**: S231-240.
Strong, L. C. (1936). "The enstablishment of the "A" strain of inbred mice." Journal of Heredity **27**(1): 21-24.
Sun, Y., W. Zhang, et al. (2014). "A glioma classification scheme based on coexpression modules of EGFR and PDGFRA." Proc Natl Acad Sci U S A **111**(9): 3538-3543.

Sutton, R. B., B. A. Davletov, et al. (1995). "Structure of the first C2 domain of synaptotagmin I: a novel Ca2+/phospholipid-binding fold." Cell **80**(6): 929-938.
Sveen, M. L., M. Schwartz, et al. (2006). "High prevalence and phenotype-genotype correlations of limb girdle muscular dystrophy type 2I in Denmark." Ann Neurol **59**(5): 808-815.
Szklarczyk, D., J. H. Morris, et al. (2017). "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible." Nucleic Acids Res **45**(D1): D362-D368.
Takahashi, T., M. Aoki, et al. (2003). "Dysferlin mutations in Japanese Miyoshi myopathy: relationship to phenotype." Neurology **60**(11): 1799-1804.
Tasca, G., F. Moro, et al. (2013). "Limb-girdle muscular dystrophy with alpha-dystroglycan deficiency and mutations in the ISPD gene." Neurology **80**(10): 963-965.
Tasca, G., M. Pescatori, et al. (2012). "Different molecular signatures in magnetic resonance imaging-staged facioscapulohumeral muscular dystrophy muscles." PLoS One **7**(6): e38779.
Taylor, C. F., N. W. Paton, et al. (2007). "The minimum information about a proteomics experiment (MIAPE)." Nat Biotechnol **25**(8): 887-893.
The UniProt Consortium (2017). "UniProt: the universal protein knowledgebase." Nucleic Acids Res **45**(D1): D158-D169.
Therrien, C., S. Di Fulvio, et al. (2009). "Characterization of lipid binding specificities of dysferlin C2 domains reveals novel interactions with phosphoinositides." Biochemistry **48**(11): 2377-2384.
Therrien, C., D. Dodig, et al. (2006). "Mutation impact on dysferlin inferred from database analysis and computer-based structural predictions." J Neurol Sci **250**(1-2): 71-78.
Thompson, R. and V. Straub (2016). "Limb-girdle muscular dystrophies - international collaborations for translational research." Nat Rev Neurol **12**(5): 294-309.
Thorley, M. (2016). Analysis of the dystrophin interactome. Paris, Berlin, UPMC, FU.
Tin Kam, H. (1998). "The random subspace method for constructing decision forests." IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(8): 832-844.
Togo, T. and R. A. Steinhardt (2004). "Nonmuscle myosin IIA and IIB have distinct functions in the exocytosis-dependent process of cell membrane repair." Mol Biol Cell **15**(2): 688-695.
Torella, A., M. Fanin, et al. (2013). "Next-generation sequencing identifies transportin 3 as the causative gene for LGMD1F." PLoS One **8**(5): e63536.
Trollet, C., T. Gidaro, et al. (1993). "Oculopharyngeal Muscular Dystrophy."
Tukey, J. W. (1977). Exploratory Data Analysis, Addison-Wesley.
Turk, R., E. Sterrenburg, et al. (2006). "Common pathological mechanisms in mouse models for muscular dystrophies." FASEB J **20**(1): 127-129.
Turner, C. and D. Hilton-Jones (2010). "The myotonic dystrophies: diagnosis and management." J Neurol Neurosurg Psychiatry **81**(4): 358-367.
Turner, S. and L. Chen (2011). "Updated Security Considerations for the MD5 Message-Digest and the HMAC-MD5 Algorithms."

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-5121.
Uaesoontrachoon, K., H. J. Cha, et al. (2013). "The effects of MyD88 deficiency on disease phenotype in dysferlin-deficient A/J mice: role of endogenous TLR ligands." J Pathol **231**(2): 199-209.
Udd, B. (2011). "Distal muscular dystrophies." Handb Clin Neurol **101**: 239-262.
Uhlen, M., L. Fagerberg, et al. (2015). "Proteomics. Tissue-based map of the human proteome." Science **347**(6220): 1260419.
Urtizberea, J. A., G. Bassez, et al. (2008). "Dysferlinopathies." Neurol India **56**(3): 289-297.
Usadel, B., T. Obayashi, et al. (2009). "Co-expression tools for plant biology: opportunities for hypothesis generation and caveats." Plant Cell Environ **32**(12): 1633-1651.
Vafiadaki, E., A. Reis, et al. (2001). "Cloning of the mouse dysferlin gene and genomic characterization of the SJL-Dysf mutation." Neuroreport **12**(3): 625-629.
Vandenbon, A., V. H. Dinh, et al. (2016). "Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system." Proc Natl Acad Sci U S A **113**(17): E2393-2402.
Velculescu, V. E., L. Zhang, et al. (1995). "Serial Analysis of Gene Expression." Science **270**(5235): 484-487.
Verhaert, D., K. Richards, et al. (2011). "Cardiac involvement in patients with muscular dystrophies: magnetic resonance imaging phenotype and genotypic considerations." Circ Cardiovasc Imaging **4**(1): 67-76.
Vernengo, L., J. Oliveira, et al. (2011). "Novel ancestral Dysferlin splicing mutation which migrated from the Iberian peninsula to South America." Neuromuscul Disord **21**(5): 328-337.
Vieira, N. M., M. S. Naslavsky, et al. (2014). "A defect in the RNA-processing protein HNRPDL causes limb-girdle muscular dystrophy 1G (LGMD1G)." Hum Mol Genet **23**(15): 4103-4110.
von der Hagen, M., S. H. Laval, et al. (2005). "The differential gene expression profiles of proximal and distal muscle groups are altered in pre-pathological dysferlin-deficient mice." Neuromuscul Disord **15**(12): 863-877.
Walt, D. R. (2000). "Techview: molecular biology. Bead-based fiber-optic arrays." Science **287**(5452): 451-452.
Walton, J. N. and F. J. Nattrass (1954). "On the classification, natural history and treatment of the myopathies." Brain **77**(2): 169-231.
Wang, J., S. Vasaikar, et al. (2017). "WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit." Nucleic Acids Res.
Wang, P., H. Qi, et al. (2015). "ImmuCo: a database of gene co-expression in immune cells." Nucleic Acids Res **43**(Database issue): D1133-1139.
Washington, N. L. and S. Ward (2006). "FER-1 regulates Ca2+ -mediated membrane fusion during C. elegans spermatogenesis." J Cell Sci **119**(Pt 12): 2552-2562.

Weiler, T., C. R. Greenberg, et al. (1996). "Limb-girdle muscular dystrophy and Miyoshi myopathy in an aboriginal Canadian kindred map to LGMD2B and segregate with the same haplotype." Am J Hum Genet **59**(4): 872-878.

Weisleder, N., H. Takeshima, et al. (2009). "Mitsugumin 53 (MG53) facilitates vesicle trafficking in striated muscle to contribute to cell membrane repair." Commun Integr Biol **2**(3): 225-226.

Weller, A. H., S. A. Magliato, et al. (1997). "Spontaneous myopathy in the SJL/J mouse: pathology and strength loss." Muscle Nerve **20**(1): 72-82.

Wenzel, K., M. Carl, et al. (2006). "Novel sequence variants in dysferlin-deficient muscular dystrophy leading to mRNA decay and possible C2-domain misfolding." Hum Mutat **27**(6): 599-600.

Wenzel, K., C. Geier, et al. (2007). "Dysfunction of dysferlin-deficient hearts." J Mol Med (Berl) **85**(11): 1203-1214.

Wenzel, K., J. Zabojszcza, et al. (2005). "Increased susceptibility to complement attack due to down-regulation of decay-accelerating factor/CD55 in dysferlin-deficient muscular dystrophy." J Immunol **175**(9): 6219-6225.

Westerhoff, H. V. and B. O. Palsson (2004). "The evolution of molecular biology into systems biology." Nat Biotechnol **22**(10): 1249-1252.

Wu, Z., R. A. Irizarry, et al. (2004). "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays." Journal of the American Statistical Association **99**(468): 909-917.

Xing, Y., K. Kapur, et al. (2006). "Probe Selection and Expression Index Computation of Affymetrix Exon Arrays." PLoS One **1**(1): e88.

Yamaji, S., A. Suzuki, et al. (2001). "A novel integrin-linked kinase-binding protein, affixin, is involved in the early stage of cell-substrate interaction." J Cell Biol **153**(6): 1251-1264.

Yang Y.H. and S. T. P. (2003). Design and analysis of comparative microarray Experiments. Statistical analysis of gene expression microarray data. S. T. P, Chapman & Hall.

Yasunaga, S., M. Grati, et al. (2000). "OTOF encodes multiple long and short isoforms: genetic evidence that the long ones underlie recessive deafness DFNB9." Am J Hum Genet **67**(3): 591-600.

Yasunaga, S., M. Grati, et al. (1999). "A mutation in OTOF, encoding otoferlin, a FER-1-like protein, causes DFNB9, a nonsyndromic form of deafness." Nat Genet **21**(4): 363-369.

Yule, G. U. (1907). "On the Theory of Correlation for any Number of Variables, Treated by a New System of Notation." Proceedings of the Royal Society of London. Series A **79**(529): 182-193.

Zhijin Wu , R. A. I., Robert Gentleman , Francisco Martinez-Murillo (2004). "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays." Forrest Spencer Journal of the American Statistical Association **99**.

Chapter 5 – Appendix

5.1 "Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd"

Published in: Nature Communications, Volume 7, 26 September 2016, Article number: 12846

Link: https://doi.org/10.1038/ncomms12846

5.2 “Changes in Communication between Muscle Stem Cells and their Environment with Aging”

Published in: Journal of Neuromuscular Diseases, Volume 2, no. 3, 2015, Pages 205-217

Link: https://doi.org/10.3233/JND-150097