Cover Page

Universiteit Leiden

Leiden University Repository

The handle http://hdl.handle.net/1887/19044 holds various files of this Leiden University dissertation.

Author: Anvar, Seyed Yahya
Title: Converging models for transcriptome studies of human diseases : the case of oculopharyngeal muscular dystrophy
Issue Date: 2012-06-06

CONVERGING MODELS for **TRANSCRIPTOME STUDIES** of **HUMAN DISEASES**

the case of **OCULOPHARYNGEAL MUSCULAR DYSTROPHY**

by

SEYED YAHYA ANVAR

Cover design and layout by: SY Anvar
Printed by: Offpage, *www.offpage.nl*

ISBN-13 978-94-6182-116-4
ISBN-10 94-6182-116-6

© Copyright 2012 by Seyed Yahya Anvar. All rights reserved. Copyright of the individual chapters rests with the authors, with the following exceptions:
Chapter 1: Oxford University Press
Chapter 2: BioMed Central Ltd.
Chapter 4: BioMed Central Ltd.
Chapter 5: Public Library of Science

No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the prior permission of the author.

Publication of this book was financially supported, in part, by the J.E. Jurriaanse Stichting.

CONVERGING MODELS FOR TRANSCRIPTOME STUDIES OF HUMAN DISEASES

THE CASE OF OCULOPHARYNGEAL MUSCULAR DYSTROPHY

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof. mr. P.F. van der Heijden

volgens besluit van het College voor Promoties

te verdedigen op woensdag 6 juni 2012

klokke 10:00 uur

door

Seyed Yahya Anvar

geboren te Tehran, Iran

in 1980

promotiecommissie

promotor:	Prof. dr. ir. Silvère M. van der Maarel
co-promotores:	dr. Peter A.C. 't Hoen
	dr. Vered Raz
	dr. Allan Tucker [1]
overige leden:	Prof. dr. Baziel G.M. van Engelen [2]
	Prof. dr. Joost N. Kok
	Prof. dr. Johan T. den Dunnen

1 Center for Intelligent Data Analysis, Brunel University, Uxbridge, United Kingdom
2 Department of Neurology, St Radboud University Medical Center, Nijmegen, The Netherlands

> “Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.”
>
> **Winston Churchill**

TABLE OF CONTENTS

11 Preface

PART ONE: Making sense out of transcriptomes, degree of similarity in diverse systems

29 **CHAPTER ONE:** Molecular and phenotypic characterization of a mouse model of oculopharyngeal muscular dystrophy reveals severe muscular atrophy restricted to fast glycolytic fibres

59 **CHAPTER TWO:** Deregulation of the ubiquitin-proteasome system is the predominant molecular pathology in OPMD animal models and patients

83 **CHAPTER THREE:** Decline in PABPN1 expression level marks skeletal muscle aging

PART TWO: Networks in biology, beyond differential expression

113 **CHAPTER FOUR:** The identification of informative genes from multiple datasets with increasing complexity

133 **CHAPTER FIVE:** Interspecies translation of disease networks increases robustness and predictive accuracy

161 Perspectives

183 Summary

185 Samenvatting

189 Abbreviations

193 Curriculum Vitae

195 Publication List

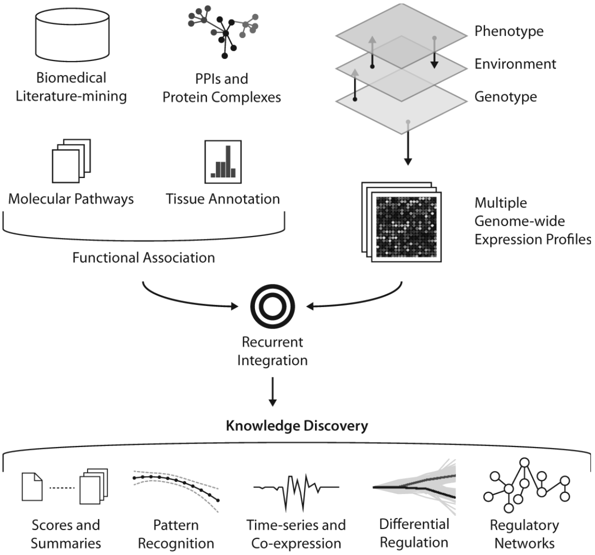
197 Acknowledgement

preface

Systems biology is the study of complex interactions between different elements of cells, tissues, and organisms. The last decade has marked the rise of systems biology owing to advancements in high-throughput techniques for genetic manipulation and measurement of cellular activities, such as genome-wide microarrays and next-generation sequencing. The advent of these technologies enabled scientists to progress beyond studying individual genes and come to a global understanding of the interplay between different elements of the cell. Despite the encouraging progress in systems biology, the high-dimensional and heterogeneous nature of biological data poses significant challenges for rigorous analysis and meaningful interpretation. For instance, differences in experimental design (such as phenotype, response, treatment, and timed events) or technical artefacts (introduced during sample preparation or data processing) complicate data integration and modelling. Notably, stochastic gene expression, even among isogenic cells, creates a source of variability at single-cell level that underlies diversified protein synthesis (Kaern et al., 2005; Kaufmann and van Oudenaarden, 2007; Ozbudak et al., 2002; Blake et al., 2003; Paulsson, 2004; Sigal et al., 2006). For instance, To and Maheshri (To and Maheshri, 2010) have shown that high or low gene expression can spontaneously be controlled by the systematic noise. This phenomenon can result from intercellular variations at the level of pathways that regulate gene expression (extrinsic noise) or arise from the random production of mRNA and bursts of protein synthesis (intrinsic noise) due to chance in interaction between cellular components. For example, genes responding to environmental stress exhibit higher level of extrinsic noise while the most robust genes regulate translation and protein degradation (Bar-Even et al., 2006; Newman et al., 2006). Thus, a full accounting of effect sizes provides crucial information on pathways and mechanisms that regulate transcriptional changes.

To tackle technical bottlenecks and arrive at biologically interpretable results, several classes of methodology have been developed, ranging from correlative approaches to those aimed to infer causal relationships. Correlation-based statistical analyses seek to identify the most prominent candidates (genes, proteins, transcription factors, or metabolites) for follow-up studies. However, the use of statistical tests that classify data points into 'changed' or 'unchanged' dismiss potentially important information on a wide range of effect sizes. Other strategies focus on the inference of modules of functionally related entities and their joint association with a biological response. Owing to the coupling and coordination of transcriptional regulation (Maniatis and Reed, 2002; Soller, 2006), rather than being independent, these modules can link the overall behaviour of a system to the interactions between its components. Thus, the use of such mathematical models can lead to the identification of prominent molecular pathways and multi-gene panels

Figure 1 – Schematic illustration of data integration. The recurrent integration of biological data (genome, transcriptome, phenome, and environment) requires special efforts in utilising data sources, protein-protein interaction (PPI) networks and protein complexes, biomedical literature, etc. The proper tuning and enhanced strategies for integrative studies leads to knowledge discovery by providing information on differential expression, the most prominent molecular pathways, common patterns, and regulatory dynamics and networks.



of interconnected regulatory networks. Nevertheless, these approaches may fail to provide mechanistic insight and discriminate between cause and consequence, which are among the main goals of systems biology. Allegorically, systems biology at its current state of development is like a group of blind wanderers studying the complex inner workings of the universe. For a panel of researchers tackling a biological question having all the tools and techniques in hand, unknown degrees of complexity make the identification of what is before them far from trivial.

To improve our methods for eliciting causal mechanisms, the use of systems with similar properties can serve as prior knowledge for benchmarking. This prior knowledge could compensate for the inherent sparseness and noisiness of high-dimensional biological data and improve precision and accuracy of their interpretation. In addition, the use of data from organisms with identical genetic background, living under controlled experimental and environmental conditions, is preferred as it results in inherently lower levels of noise and stochasticity. Integration of data from a number of model organisms may, therefore, advance the understanding of more complex biological systems. The development of strategies for robust translation of findings from one organism to another constitutes the core of this thesis. In this introduction, I outline alternative methods for inference of biologically relevant relationships, ranging from simple searches in biological modules to data-mining, machine learning, and modelling of Bayesian networks.

Data integration

Data integration consists of efforts in combining multiple datasets to provide a unified view of biological information. There is a necessity for data-mining that goes beyond the analysis of individual datasets. Hence, consensus and precision in biological interpretation can be reached only through another source of information (Tenenbaum et al., 2011). Integration of data and genomic information from multiple experiments can ultimately provide significant mechanistic insights on genomic, transcriptomic, proteomics, metabolomics, and epigenomic changes that give rise to specific phenotypes at the molecular, cellular, or organismal level (**Figure 1**). Nonetheless, the process of data integration requires a fine tuning and vigorous setting for optimal precision of findings. Various data integration strategies, at different levels, can potentially offer different views on the same biological information. High-level integration methodologies, such as meta-analyses, are dependent on filtering protocols (i.e. selection of differentially expressed genes as input) with basic assumptions which can lead to loss of biological information. Nevertheless, these approaches are useful for obtaining a gross overview over the data (Ficenec et al., 2003). In contrast, low-level approaches, such as data-mining, can facilitate the use of mutual information to gain better power in retrieving valuable information (Choi et al., 2004). Multi-layer integration of biological data may offer the best of both strategies. This approach provides a framework in which the influence of platform- or experiment-specific noise (Aitchison and Galitski, 2003) can be reduced since it reinforces the mutual information standing out above uncorrelated noise (Choi et al., 2004; Jiang et al., 2004).

Ups and downs at the transcriptome

The work presented in this thesis is largely confined to transcriptome data analyses. The amount of mRNA in the cell is finely regulated in a spatial-temporal manner to ensure cellular homeostasis. The centrality of RNA processing (Sharp, 2009), together with the comprehensive nature of current RNA expression profiling approaches, makes transcriptome data ideal for modelling of biological responses. Nevertheless, transcriptome analyses disregard important levels of regulation at the translational and post-translational level. Recent studies have demonstrated rather poor correlations between mRNA and protein levels (Guo et al., 2010; Selbach et al., 2008).

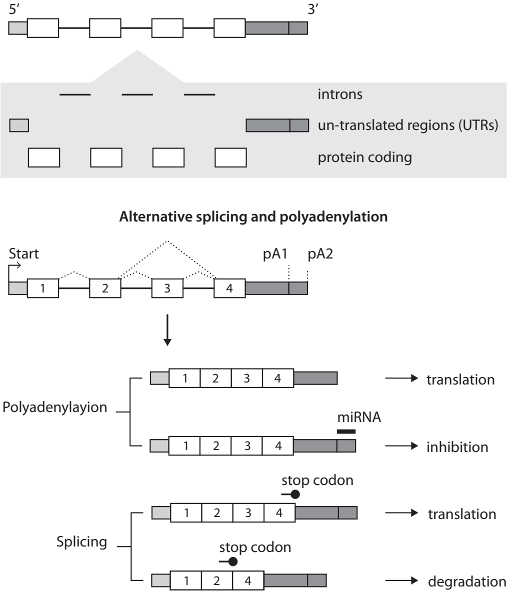


Figure 2 – Schematic overview of RNA processing and its regulation. A single gene can generate pre-mRNAs that are alternatively processed to generate a diverse set of mature mRNAs. These isoforms can differ in inclusion of exons (alternative splicing) and the polyadenylation sites in the 3’ UTR (alternative polyadenylation). Alternative protein-coding regions are depicted as mutually exclusive splicing of the third exon and selection of one of the two possible poly(A) sites (pA1 and pA2). Alternative splicing, for instance, can lead to coding frame-shifts which results in degradation of mRNA by nonsense-mediated decay pathway. On the other hand, elongation of the 3’ UTR can alter the range of regulatory elements such as microRNAs (miRNA) targeting the transcript to be subjected to different forms of post-transcriptional regulation, in this case inhibition. Additional events, such as selection of alternative first exons, can further diversify the pool of mRNAs.

The study of the transcriptome, in particular that of higher eukaryotes, is complicated by extensive RNA processing steps which give rise to different transcript variants. RNA processing events, such as splicing (Cooper et al., 2009; Wahl et al., 2009), polyadenylation (Lutz, 2008), RNA editing (Bass, 2002; Wulff et al., 2011) and other post-transcriptional modifications, widely expand the mRNA pool and, therefore, coding of an even more diverse set of functional proteins and RNA species (**Figure 2**). These events are vital for many physiological and pathophysiological processes. This may explain some of the relatively diverse phenotypic characteristics of human and chimpanzee that share 99.7% identical sequence in genome-coding regions (Calarco et al., 2007). In humans, more than 90% of genes are alternatively spliced in a tissue and cell-specific manner (Wang et al., 2008a). Like regulation of transcription, post-transcriptional processes are tightly controlled. For instance, there is an important regulatory role for microRNAs on mRNA stability and translational efficacy (Filipowicz et al., 2008) and epigenetic changes mediated by non-coding RNAs (Wang et al., 2008b; Cam et al., 2009). The integrity of these processes are controlled by mRNA stability and turnover machineries (Houseley and Tollervey, 2009) as abnormal RNA processing can lead to futile or ultimately lethal function of encoded protein. Hence, the study of transcriptional and post-transcriptional control of mRNA expression is essential for a better understanding of physiology and pathophysiology. Furthermore, the comparison of transcriptome profiles from different cell types and organisms can help determining the frequency of alternative processes and the extent to which it is subjected to species- or tissue-specific regulation (Licatalosi and Darnell, 2010).

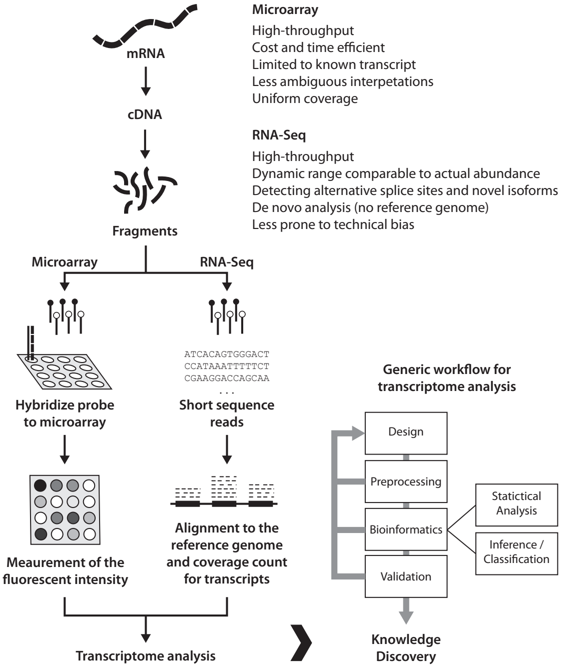

mRNA
cDNA
Fragments
Microarray
RNA-Seq
Hybridize probe to microarray
ATCACAGTGGGACT
CCATAAATTTTTCT
CGAAGGACCAGCAA
...
Short sequence reads
Meaurement of the fluorescent intensity
Alignment to the reference genome and coverage count for transcripts
Transcriptome analysis
Microarray
High-throughput
Cost and time efficient
Limited to known transcript
Less ambiguous interpetations
Uniform coverage
RNA-Seq
High-throughput
Dynamic range comparable to actual abundance
Detecting alternative splice sites and novel isoforms
De novo analysis (no reference genome)
Less prone to technical bias
Generic workflow for transcriptome analysis
Design
Preprocessing
Bioinformatics
Validation
Statictical Analysis
Inference / Classification
Knowledge Discovery

Figure 3 – Workflows for transcriptome analysis. Microarray and RNA-Seq are the most common high-throughput techniques for transcriptome profiling. The main characteristics of microarrays and RNA-Seq for transcriptome studies are listed. The general pipeline for conducting a transcriptome study involves recurring steps of experimental design, data processing, statistical analysis and network inference, and the validation of findings.

Genome-wide expression microarrays and RNA-Seq (next-generation RNA sequencing) are currently the most important technologies for transcriptome profiling (**Figure 3**). Microarrays have become one of the most commonly used tools in transcriptomics studies owing to their cost-efficiency and speed in simultaneously measuring thousands of gene transcripts. In addition, microarrays have been designed with distinct features to address the RNA complexity such as exon-junction arrays for capturing differential splicing events (Johnson et al., 2003). Despite their obvious potency, microarrays are limited by gene annotations and can only detect known transcripts for which microarray probes have been designed, whilst novel transcripts and transcript variants will be missed. Moreover, the technical noise in microarray signals, being dependent on probe hybridization and annealing properties, is relatively high. This negatively affects data reproducibility and cross-platform and sample comparisons (Ioannidis et al., 2009). RNA-Seq, on the other hand, generates millions of reads and has the potential to measure the complete transcriptome including alternative splicing and polyadenylation, and RNA editing events (Pan et al., 2008; Wang et al., 2008a). Nevertheless, RNA-Seq analysis strategies are currently under development as exact quantification of the relative abundance of different transcript variants remains challenging.

Rewiring regulatory networks in biology

Biological processes do not occur by isolated genes or proteins but act through functional regulatory networks. The degree to which gene products appear in the cell and exert their function is regulated by such biological networks. Therefore, the implications of gene defects would not be restricted to the activity of specific gene products but can have many severe effects by spreading along sub-network structures (Barabasi et al., 2011). This interconnectivity implies that the identification of regulatory networks and understanding the evolution and structural features

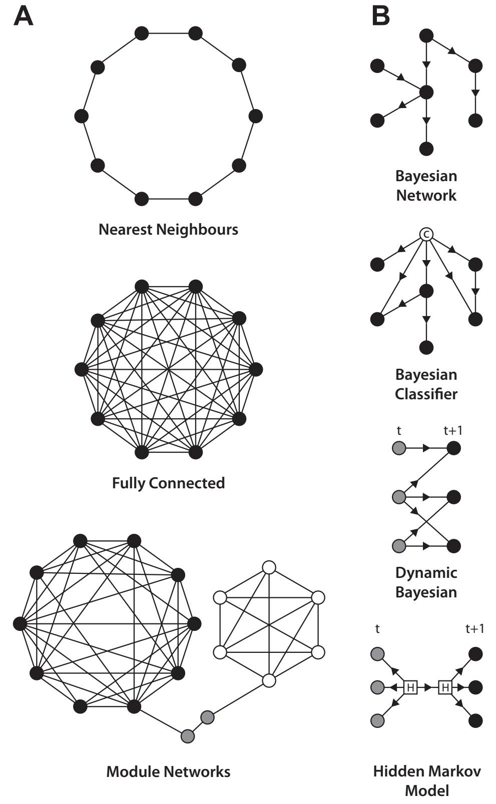


Figure 4 – Schematic illustration of biological networks. A) Co-expression networks can be constructed under various constraints and settings. A cluster of ten nodes can be interconnected on the basis of their nearest neighbours, depicted as a ring. Fully connected networks of ten nodes represent a cluster of fully interconnected nodes where all nodes are co-expressed. Co-expression networks can also be represented as connected modules. Here, a cluster of ten partially connected nodes (black) are linked to a cluster of six partially connected nodes (white) through two independent nodes (gray). **B)** A Bayesian network that encodes a joint distribution is very flexible and can be constructed in different architectures based upon the data analysis task: Bayesian networks, Bayesian classifiers (these networks include a class node, depicted by C, for prediction), dynamic Bayesian networks (these networks support time-series where nodes represent variables at a point in time), and hidden Markov model (these networks can handle unmeasured information by incorporating a hidden or latent variable, depicted by H).

of specific networks are vital for better understanding the phenotypic impacts of genetic defects and the associated complications (Schadt, 2009; Goldstein, 2009; Karlebach and Shamir, 2008). Thus, as genetics is aimed to answer the question of 'what', network-based models are designed to go one step beyond by tackling the question of 'how'.

Network-based approaches have transformed the field of systems biology. These approaches are mainly expression-centric and can be classified into two types of module inference and transcription regulatory network. The first type of analysis involves the study of co-expression networks (**Figure 4A**). This comprises the identification of functional relationships between genes under the assumption that genes with similar function exhibit interrelated expression patterns and can be described as a functional module (Stuart et al., 2003). These methods require careful interpretation as they are highly sensitive to noise. Such models are biased towards identification of relationships between genes that are tightly co-expressed and disregard those that do not exhibit sufficient co-expression profiles with other genes (Michoel et al., 2009). It is important to bear in mind that correlation does not imply causation. This issue can be partially addressed by the use of time-series data. In the second type of approach, methods go one step further by taking into account the sense of similarity, representativeness, and randomness of biological data. These models can accommodate hidden variables, assess the causality of relationships and, most importantly, provide reasoning and predictions for unseen data (**Figure 4B**). Nonetheless, these models are prone to overfitting and generation of multiple probable solutions that can be circumvented by the use of multiple independent datasets.

The use of prior knowledge about functional interactions has been shown to successfully reduce the search space and to make networks more robust (Segal et al., 2003; Pe'er et al., 2002; Steele et al., 2009). This method works for well-studied diseases or biological systems, but is less likely to identify novel regulatory interactions that are involved in the underlying molecular mechanisms of rare or complex disorders. In addition, this bias can falsely expose the network to sample differences in the absence of a biological cause. In this thesis, the unbiased use of independent datasets from different organisms as prior knowledge is further explored (**Figure 5A**). Modular structure of regulatory networks (Ma et al., 2004) and largely conserved functional properties of genes across species provide a detailed framework for identification of relationships that are conserved across species. It was hypothesized that relationships that are identified in an interspecies gene network are also biologically more meaningful. Furthermore, they result in more reliable identification of key players in biological processes under study. However, translation of regulatory networks across different platforms or organisms is far from trivial. This is evident from our limited knowledge of true protein orthologues and transcript variants coding for proteins with similar functions in different species. For this, new algorithms and optimization techniques needed to be developed (Chapter four and five).

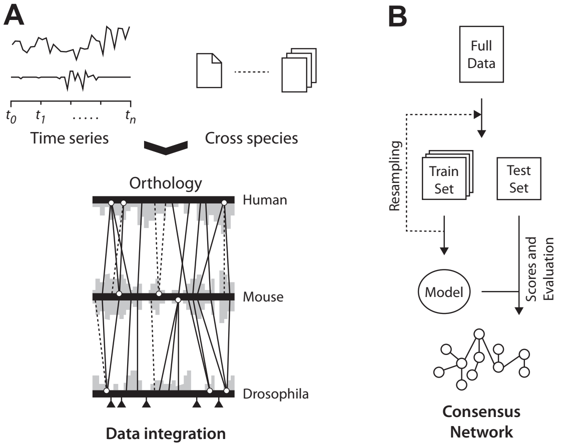


Figure 5 – Bayesian regulatory networks in computational biology. A) Interspecies (or inter-platform) integration can be achieved by taking into account the many-to-many relationships of orthologue genes/transcripts (depicted by circles). Depending on the technology used for generating biological data, the information and coverage on the possible orthologues and their transcripts varies (depicted in gray). **B)** The process of building a prediction model involves partitioning data into training and test folds at random. Next, after constructing and tuning the parameters, models are tested on the test data. This process is repeated by resampling from the full data until all partitions are used for building and testing the models. The consensus network can be reached by averaging and assessing all the constructed models. A number of different computational techniques can be used to optimise the partitioning, building, and averaging these networks. The consensus model, the key nodes, and the predictions can reveal new biological insights.

Among the possible approaches for modelling of biological networks, Bayesian networks have certain advantages as they are able to deal with uncertainties and stochastic effects (Pearl, 1988; Friedman, 2004; Friedman et al., 2000; Segal et al., 2003). A Bayesian network can encode gene interaction by modelling the joint probability distribution that represents possible transcriptional behaviour for a set of genes. It consists of a directed acyclic graph (DAG) that denotes conditional independencies and a conditional probability distribution for each gene (represented by a node in the graph). These networks can represent complex relationships between genes and are capable of integrating different types of data (from phenotypic and genotypic categorical data to continuous gene expression profiles). In addition, the probabilistic nature of such networks can easily accommodate noise or missing data by weighting each information source according to its reliability. In contrast to many statistical models, the transparent nature of Bayesian networks (in terms of the graphical structure and local probability distributions) leads to better interpretation and understanding of the underlying biological processes. The combination of a rigorous training and testing regime (including cross-validation which is a statistical method for assessing the performance of a fitted model in predicting the observation made on unseen data) and optimization procedures (such as simulated annealing) can lead to the inference of reliable network structure (**Figure 5B**).

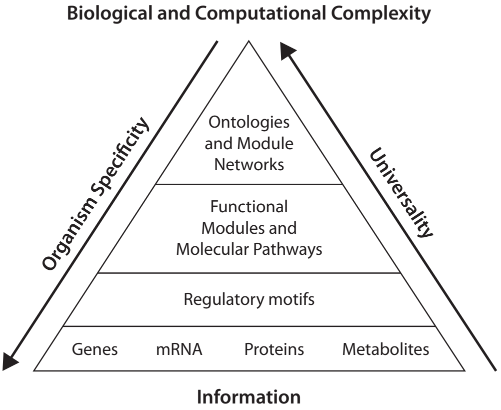

Biological and Computational Complexity
Ontologies and Module Networks
Functional Modules and Molecular Pathways
Regulatory motifs
Genes
mRNA
Proteins
Metabolites
Information
Organism Specificity
Universality

Figure 6 – Complexity pyramid, from individual to mutual. The bottom of the pyramid represents the functional components of the cell for which high-throughput biological data are produced (level 1). The next layer brings complex regulatory motifs (level 2) function in a highly spatial-temporal manner to provide diverse sets of functional modules. These sub-networks are the building blocks of molecular pathways (level 3). Modules of functionally related entities work as components of a nested structure that represents context-oriented global organisation of living organisms. Although the individual elements of these networks can be unique to a given organism, the topologic properties of module networks share a high degree of similarities.

Model systems and the study of human diseases

Biomedical research has evolved around model organisms which have played a central role in the studies of human disorders. In spite of growing achievements in genome-wide association studies and whole-genome profiling, genetic studies of human diseases are significantly limited owing to factors such as environmental influences and genetic heterogeneity. The challenges posed by human genetic research can potentially be circumvented in model organisms. This is due to much simplified and experimentally traceable system that provides unbiased environment for characterization of genetic data (Aitman et al., 2011). Nevertheless, model systems have their own limitations and cannot fully replace the human data as genetic architecture and complex traits, such as epigenetic and environmental effects, are hard to replicate in model organisms. Moreover, genetic engineering may introduce significant artefacts. Thus, data from model organisms should be interpreted with care. In addition, the use of multiple model organisms may be necessary to identify the most prominent and disease-related molecular mechanisms that can be projected on human data with high precision. The design of such integrative strategies would bridge the gap between less noisy data from model systems to more stochastic human biology.

As model systems, along with high-throughput transcriptional profiling, continue to transform the study of human disorders, novel algorithms are needed to capture, characterize, and model the hierarchy and dynamics of biological data (**Figure 6**). It is clear that attentive modelling and optimization of integration strategy would ultimately serve as a powerful system for knowledge discovery in the study of human genetic disorders.

Oculopharyngeal muscular dystrophy

In this study, I have focused my efforts on the improved understanding of disease mechanisms in oculopharyngeal muscular dystrophy (OPMD). OPMD is an autosomal dominant and late-onset disorder, usually manifest in midlife (after the age of 40). OPMD symptoms are progressive and characterised by *ptosis*, *dysphagia*, and weakness of proximal limb (**Figure** 7). As the disease progresses, muscle weakness can spread to additional skeletal muscles such as facial muscle weakness, tongue atrophy, and dysphonia (Brais and Rouleau, 1993). In some OPMD patients, reports have indicated mental retardation, cognitive impairment, spinal cord involvement, and dementia as additional symptoms (Millefiorini and Filippini, 1967; Sarkar et al., 1995; Blumen et al., 2009; Linoli et al., 1991; Mizoi et al., 2011; Dubbioso et al., 2011). In spite of these observations, the main OPMD symptoms are restricted to voluntary muscles. However, the degree to which these muscles are affected and the associated age of onset is variable. Nevertheless, by the time the dis-

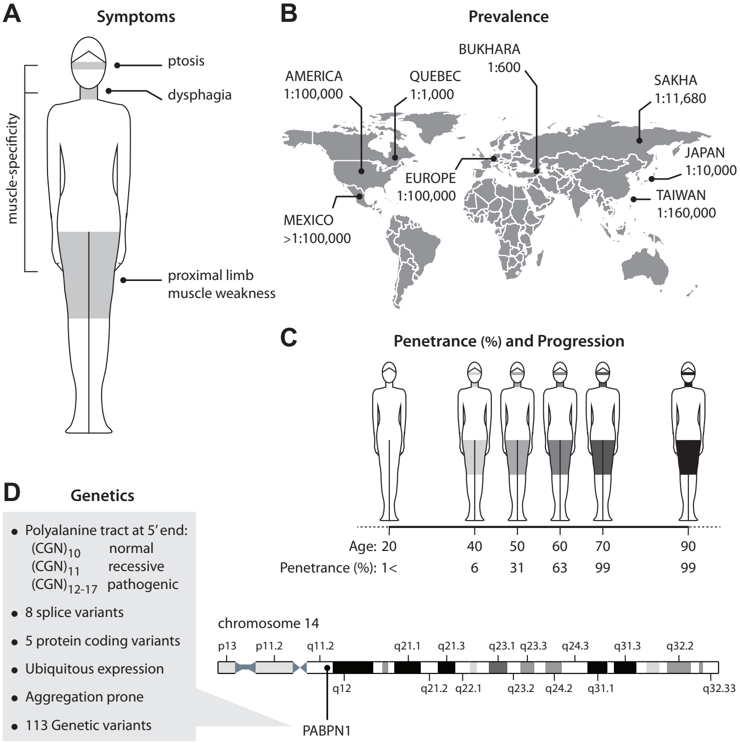


Figure 7 – Schematic characterisation of oculopharyngeal muscular dystrophy. A) OPMD symptoms are mainly restricted to skeletal muscles. **B)** Divers prevalence rates of OPMD estimated in different populations. Worldwide prevalence is estimated to be 1:100,000. **C)** Penetrance (%) and progression rate of OPMD is depicted. **D)** Overview of genetic information for the PABPN1 and pathogenic mutations.

ease is progressed, the quality of life is greatly affected as *ptosis* can cause visual limitations, *dysphagia* may lead to aspiration pneumonia and weight loss, and patients with proximal limb weakness can eventually be wheelchair bound. OPMD is a rare disorder with estimated prevalence of 1 in 100,000 in western countries (Fan and Rouleau, 2003). However, there is a vast diversity of prevalence between different populations (Pulkes et al., 2011; Brais and Rouleau, 1993; Semmler et al., 2007; Uyama et al., 1997; Maksimova et al., 2007; Puzyrev and Maximova, 2008; Agarwal et al., 2012). In some isolated populations the incidence is much higher, among which the Bukhara originated Jewish community (1 in 600) and French-Canadian populations (1 in 1000) have the highest prevalence (Brais et al., 1995; Blumen et al., 1997).

OPMD is caused by expansion of a homopolymeric alanine (Ala) stretch at the N-terminus of the Poly(A) Binding Protein Nuclear 1 (PABPN1) (Brais et al., 1998). While wild-type *PABPN1* contains a $(GCN)_{10}$ repeat within the first exon, in the mutated form it holds an expanded repeat of $(CGN)_{12\text{-}17}$ that leads to 2-7 additional Ala residues. The most frequently occurring mutation is estimated to be the expansion of the GCG from 6 to 9 repeats whilst other mutations (such as the combination of GCA and GCG expansions) have also been reported (Nakamoto et al., 2002; Scacheri et al., 1999; Robinson et al., 2006). The *PABPN1* gene is located on chromosome 14q11.2 and has 8 splice variants, 5 of which encode functional proteins (**Figure 7**). The encoded protein localizes mostly in the nucleus and to a lower extent in the cytoplasm. Within the nucleus, PABPN1 is enriched in nuclear speckles (subnuclear structures that are enriched in pre-mRNA and are located in interchromatic regions). Wide-type PABPN1 has multiple roles in mRNA processing, stability and translation, among which the role of PABPN1 in mRNA polyadenylation has been extensively studied (Kuhn et al., 2009; Wahle, 1991; Wahle, 1995; Apponi et al., 2010). PABPN1 protein is also involved in the export of mRNAs from the nucleus to the cytoplasm (Apponi et al., 2010; Calado et al., 2000a; Brune et al., 2005).

The underlying molecular mechanisms by which the mutated PABPN1 causes progressive muscle weakness are not fully understood. In spite of the ubiquitous expression of PABPN1, the clinical and pathological features of OPMD are initially restricted to a subset of skeletal muscles. The wild-type and expanded PABPN1 (expPABPN1) are prone to aggregation (David et al., 2010; Klein et al., 2008). PABPN1 accumulates in intranuclear inclusions (INI) in 1-3% of myonuclei (Tome and Fardeau, 1980; Calado et al., 2000b). To better understand the molecular mechanisms leading to OPMD, animal models for OPMD were generated in *Drosophila,* mouse and *C. elegans* with high overexpression of expPABPN1 under a muscle-specific promoter (Chartier et al., 2006; Davies et al., 2005; Catoire et al., 2008). These model systems recapitulate INI formation and progressive muscle weakness observed in OPMD. A correlation between INI formation and muscle weakness has been reported in these models (Chartier et al., 2006; Davies et al., 2005; Catoire et al., 2008). In addition, it has been shown that protein disaggregation approaches can attenuate muscle symptoms in OPMD model systems (Davies et al., 2006; Catoire et al., 2008; Chartier et al., 2009). Nevertheless, in a mouse model with low overexpression of expPABPN1, muscle symptoms were not observed (Hino et al., 2004). Naturally occurring wild-type PABPN1 inclusions with fibril structures have also been reported in oxytocin-producing neurons (Berciano et al., 2004; Villagra et al., 2008). In contrast to INI formation in OPMD, the inclusions of wild-type PABPN1 do not cause a disease. Differing transitional pre-inclusion foci and structural characteristics have been shown between the wild-type and expanded PABPN1 (Raz et al., 2011). Therefore, differences in processes that precede the formation of INIs suggest the cytotoxic structure of the pre-aggregated proteins.

The complexity of the underlying mechanisms and the low prevalence of OPMD call for multidisciplinary and combined efforts to decipher disease mechanisms. As the focus of the current thesis, exhaustive use of the state-of-the-art data-mining strategies and cross-species data integration can provide a comprehensive, less technically biased, and more accurate mechanistic insights on the disease pathogenesis. Understanding the underlying causes of OPMD is a key step toward enabling earlier and more precise diagnosis, prognosis, therapeutic interventions, and drug discovery.

Thesis overview

In this thesis, I have mainly focused on interdisciplinary approaches for biomedical knowledge discovery. This required special efforts in developing systematic strategies to integrate various data sources and techniques, leading to improved discovery of mechanistic insights of human diseases. Chapter **one** looks at the possibility in which combining various bioinformatics-based strategies can significantly improve the characterization of the OPMD mouse model. We discuss that this approach in knowledge discovery, on the basis of our extensive analysis, helped us to shed some light on how this model system relates to OPMD pathophysiology in human. In Chapter **two**, we expand on this combinatory approach by conducting a cross-species data analysis. In this study, we have looked for common patterns that emerge by assessing the transcriptome data from three OPMD model systems and patients. This strategy led to unravelling the most prominent molecular pathway involved in OPMD pathology. The **third** Chapter achieves a similar goal to identify similar molecular and pathophysiological features between OPMD and the common process of skeletal muscle ageing. Engaging in a study in which the focus was made on the universality of biological processes, in the light of evolutionary mechanisms and common functional features, led to novel discoveries. This work helped us to uncover remarkable insights on molecular mechanisms of ageing muscles and protein aggregation. Chapters **four** and **five** take a different route by tackling the field of computational biology. These chapters aim to extend network inference by providing novel strategies for the exploitation and integration of multiple data sources. We show that these developments allow us to infer more robust regulatory mechanisms to be identified while translations and predictions are made across very different datasets, platforms, and organisms. Finally, I close this thesis by providing an outlook on ways the field of systems biology can evolve in order to offer enhanced, diversified and robust strategies for knowledge discovery.

Reference List

Agarwal,P.K., Mansfield,D.C., Mechan,D., Al-Shahi,S.R., Davenport,R.J., Connor,M., Metcalfe,R., and Petty,R. (2012). Delayed diagnosis of oculopharyngeal muscular dystrophy in Scotland. Br. J. Ophthalmol. *96*, 281-283.

Aitchison,J.D. and Galitski,T. (2003). Inventories to insights. J. Cell Biol. *161*, 465-469.

Aitman,T.J., Boone,C., Churchill,G.A., Hengartner,M.O., Mackay,T.F., and Stemple,D.L. (2011). The future of model organisms in human disease research. Nat Rev Genet. 12, 575-582.

Apponi,L.H., Leung,S.W., Williams,K.R., Valentini,S.R., Corbett,A.H., and Pavlath,G.K. (2010). Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. Hum. Mol. Genet. *19*, 1058-1065.

Bar-Even,A., Paulsson,J., Maheshri,N., Carmi,M., O'Shea,E., Pilpel,Y., and Barkai,N. (2006). Noise in protein expression scales with natural protein abundance. Nat Genet. *38*, 636-643.

Barabasi,A.L., Gulbahce,N., and Loscalzo,J. (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet. *12*, 56-68.

Bass,B.L. (2002). RNA editing by adenosine deaminases that act on RNA. Annu. Rev Biochem. *71*, 817-846.

Berciano,M.T., Villagra,N.T., Ojeda,J.L., Navascues,J., Gomes,A., Lafarga,M., and Carmo-Fonseca,M. (2004). Oculopharyngeal muscular dystrophy-like nuclear inclusions are present in normal magnocellular neurosecretory neurons of the hypothalamus. Hum. Mol. Genet *13*, 829-838.

Blake,W.J., KAErn,M., Cantor,C.R., and Collins,J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633-637.

Blumen,S.C., Bouchard,J.P., Brais,B., Carasso,R.L., Paleacu,D., Drory,V.E., Chantal,S., Blumen,N., and Braverman,I. (2009). Cognitive impairment and reduced life span of oculopharyngeal muscular dystrophy homozygotes. Neurology *73*, 596-601.

Blumen,S.C., Nisipeanu,P., Sadeh,M., Asherov,A., Blumen,N., Wirguin,Y., Khilkevich,O., Carasso,R.L., and Korczyn,A.D. (1997). Epidemiology and inheritance of oculopharyngeal muscular dystrophy in Israel. Neuromuscul. Disord. *7 Suppl 1*, S38-S40.

Brais,B., Bouchard,J.P., Xie,Y.G., Rochefort,D.L., Chretien,N., Tome,F.M., Lafreniere,R.G., Rommens,J.M., Uyama,E., Nohira,O., Blumen,S., Korczyn,A.D., Heutink,P., Mathieu,J., Duranceau,A., Codere,F., Fardeau,M., and Rouleau,G.A. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet *18*, 164-167.

Brais,B. and Rouleau,G.A. (1993). Oculopharyngeal Muscular Dystrophy.

Brais,B., Xie,Y.G., Sanson,M., Morgan,K., Weissenbach,J., Korczyn,A.D., Blumen,S.C., Fardeau,M., Tome,F.M., Bouchard,J.P., and . (1995). The oculopharyngeal muscular dystrophy locus maps to the region of the cardiac alpha and beta myosin heavy chain genes on chromosome 14q11.2-q13. Hum. Mol. Genet *4*, 429-434.

Brune,C., Munchel,S.E., Fischer,N., Podtelejnikov,A.V., and Weis,K. (2005). Yeast poly(A)-binding protein Pab1 shuttles between the nucleus and the cytoplasm and functions in mRNA export. RNA. *11*, 517-531.

Calado,A., Kutay,U., Kuhn,U., Wahle,E., and Carmo-Fonseca,M. (2000a). Deciphering the cellular pathway for transport of poly(A)-binding protein II. RNA. *6*, 245-256.

Calado,A., Tome,F.M., Brais,B., Rouleau,G.A., Kuhn,U., Wahle,E., and Carmo-Fonseca,M. (2000b). Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. Hum. Mol. Genet *9*, 2321-2328.

Calarco,J.A., Xing,Y., Caceres,M., Calarco,J.P., Xiao,X., Pan,Q., Lee,C., Preuss,T.M., and Blencowe,B.J. (2007). Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev. *21*, 2963-2975.

Cam,H.P., Chen,E.S., and Grewal,S.I. (2009). Transcriptional scaffolds for heterochromatin assembly. Cell *136*, 610-614.

Catoire,H., Pasco,M.Y., Abu-Baker,A., Holbert,S., Tourette,C., Brais,B., Rouleau,G.A., Parker,J.A., and Neri,C. (2008). Sirtuin inhibition protects from the polyalanine muscular dystrophy protein PABPN1. Hum. Mol. Genet *17*, 2108-2117.

Chartier,A., Benoit,B., and Simonelig,M. (2006). A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. EMBO J *25*, 2253-2262.

Chartier,A., Raz,V., Sterrenburg,E., Verrips,C.T., van der Maarel,S.M., and Simonelig,M. (2009). Prevention of oculopharyngeal muscular dystrophy by muscular expression of Llama single-chain intrabodies in vivo. Hum. Mol. Genet.

Choi,J.K., Choi,J.Y., Kim,D.G., Choi,D.W., Kim,B.Y., Lee,K.H., Yeom,Y.I., Yoo,H.S., Yoo,O.J., and Kim,S. (2004). Integrative analysis of multiple gene expression profiles applied to liver cancer study. FEBS Lett. *565*, 93-100.

Cooper,T.A., Wan,L., and Dreyfuss,G. (2009). RNA and disease. Cell *136*, 777-793.

David,D.C., Ollikainen,N., Trinidad,J.C., Cary,M.P., Burlingame,A.L., and Kenyon,C. (2010). Widespread protein aggregation as an inherent part of aging in C. elegans. PLoS. Biol. *8*, e1000450.

Davies,J.E., Sarkar,S., and Rubinsztein,D.C. (2006). Trehalose reduces aggregate formation and delays pathology in a transgenic mouse model of oculopharyngeal muscular dystrophy. Hum. Mol. Genet *15*, 23-31.

Davies,J.E., Wang,L., Garcia-Oroz,L., Cook,L.J., Vacher,C., O'Donovan,D.G., and Rubinsztein,D.C. (2005). Doxycycline attenuates and delays toxicity of the oculopharyngeal muscular dystrophy mutation in transgenic mice. Nat Med. *11*, 672-677.

Dubbioso,R., Moretta,P., Manganelli,F., Fiorillo,C., Iodice,R., Trojano,L., and Santoro,L. (2011). Executive functions are impaired in heterozygote patients with oculopharyngeal muscular dystrophy. J. Neurol.

Fan,X. and Rouleau,G.A. (2003). Progress in understanding the pathogenesis of oculopharyngeal muscular dystrophy. Can. J. Neurol. Sci. *30*, 8-14.

Ficenec,D., Osborne,M., Pradines,J., Richards,D., Felciano,R., Cho,R.J., Chen,R.O., Liefeld,T., Owen,J., Ruttenberg,A., Reich,C., Horvath,J., and Clark,T. (2003). Computational knowledge integration in biopharmaceutical research. Brief. Bioinform. *4*, 260-278.

Filipowicz,W., Bhattacharyya,S.N., and Sonenberg,N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet. *9*, 102-114.

Friedman,N. (2004). Inferring cellular networks using probabilistic graphical models. Science *303*, 799-805.

Friedman,N., Linial,M., Nachman,I., and Pe'er,D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601-620.

Goldstein,D.B. (2009). Common genetic variation and human traits. N. Engl. J. Med. *360*, 1696-1698.

Guo,H., Ingolia,N.T., Weissman,J.S., and Bartel,D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835-840.

Hino,H., Araki,K., Uyama,E., Takeya,M., Araki,M., Yoshinobu,K., Miike,K., Kawazoe,Y., Maeda,Y., Uchino,M., and Yamamura,K. (2004). Myopathy phenotype in transgenic mice expressing mutated PABPN1 as a model of oculopharyngeal muscular dystrophy. Hum. Mol. Genet *13*, 181-190.

Houseley,J. and Tollervey,D. (2009). The many pathways of RNA degradation. Cell *136*, 763-776.

Ioannidis,J.P., Allison,D.B., Ball,C.A., Coulibaly,I., Cui,X., Culhane,A.C., Falchi,M., Furlanello,C., Game,L., Jurman,G., Mangion,J., Mehta,T., Nitzberg,M., Page,G.P., Petretto,E., and van,N., V (2009). Repeatability of published microarray gene expression analyses. Nat Genet. *41*, 149-155.

Jiang,H., Deng,Y., Chen,H.S., Tao,L., Sha,Q., Chen,J., Tsai,C.J., and Zhang,S. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC. Bioinformatics. *5*, 81.

Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R., and Shoemaker,D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.

KAErn,M., Elston,T.C., Blake,W.J., and Collins,J.J. (2005). Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet. *6*, 451-464.

Karlebach,G. and Shamir,R. (2008). Modelling and analysis of gene regulatory networks. Nat Rev Mol. Cell Biol. *9*, 770-780.

Kaufmann,B.B. and van Oudenaarden,A. (2007). Stochastic gene expression: from single molecules to the proteome. Curr. Opin. Genet. Dev. *17*, 107-112.

Klein,A.F., Ebihara,M., Alexander,C., Dicaire,M.J., Sasseville,A.M., Langelier,Y., Rouleau,G.A., and Brais,B. (2008). PABPN1 polyalanine tract deletion and long expansions modify its aggregation pattern and expression. Exp. Cell Res. *314*, 1652-1666.

Kuhn,U., Gundel,M., Knoth,A., Kerwitz,Y., Rudel,S., and Wahle,E. (2009). Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J. Biol. Chem. *284*, 22803-22814.

Licatalosi,D.D. and Darnell,R.B. (2010). RNA processing and its regulation: global insights into biological networks. Nat Rev Genet. *11*, 75-87.

Linoli,G., Tomelleri,G., and Ghezzi,M. (1991). Oculopharyngeal muscular dystrophy. Description of a case with involvement of the central nervous system]. Pathologica *83*, 325-334.

Lutz,C.S. (2008). Alternative polyadenylation: a twist on mRNA 3' end formation. ACS Chem. Biol. *3*, 609-617.

Ma,H.W., Buer,J., and Zeng,A.P. (2004). Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC. Bioinformatics. *5*, 199.

Maksimova,N.R., Korotov,M.N., and Nikolaeva,I.A. (2007). Clinical and molecular genetic aspects of Oculopharyngeal Muscular Dystrophy in Republic of Sakha (Yakutiya). Genetika i patologiya 160-161.

Maniatis,T. and Reed,R. (2002). An extensive network of coupling among gene expression machines. Nature *416*, 499-506.

Michoel,T., De,S.R., Joshi,A., Van de Peer,Y., and Marchal,K. (2009). Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. BMC. Syst. Biol. *3*, 49.

Millefiorini,M. and Filippini,C. (1967). Oculopharyngeal muscular dystrophy. Riv. Neurol. *37*, 327-337.

Mizoi,Y., Yamamoto,T., Minami,N., Ohkuma,A., Nonaka,I., Nishino,I., Tamura,N., Amano,T., and Araki,N. (2011). Oculopharyngeal muscular dystrophy associated with dementia. Intern. Med. *50*, 2409-2412.

Nakamoto,M., Nakano,S., Kawashima,S., Ihara,M., Nishimura,Y., Shinde,A., and Kakizuka,A. (2002). Unequal crossing-over in unique PABP2 mutations in Japanese patients: a possible cause of oculopharyngeal muscular dystrophy. Arch. Neurol. *59*, 474-477.

Newman,J.R., Ghaemmaghami,S., Ihmels,J., Breslow,D.K., Noble,M., DeRisi,J.L., and Weissman,J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840-846.

Ozbudak,E.M., Thattai,M., Kurtser,I., Grossman,A.D., and van,O.A. (2002). Regulation of noise in the expression of a single gene. Nat Genet. *31*, 69-73.

Pan,Q., Shai,O., Lee,L.J., Frey,B.J., and Blencowe,B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. *40*, 1413-1415.

Paulsson,J. (2004). Summing up the noise in gene networks. Nature *427*, 415-418.

Pe'er,D., Regev,A., and Tanay,A. (2002). Minreg: inferring an active regulator set. Bioinformatics. *18 Suppl 1*, S258-S267.

Pearl,J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. (San Francisco: Morgan Kaufmann).

Pulkes,T., Papsing,C., Busabaratana,M., Dejthevaporn,C., and Witoonpanich,R. (2011). Mutation and haplotype analysis of oculopharyngeal muscular dystrophy in Thai patients. J. Clin. Neurosci. *18*, 674-677.

Puzyrev,V.P. and Maximova,N.P. (2008). Hereditary Diseases among Yakuts. Russian Journal of Genetics *44*, 1141-1147.

Raz,V., Abraham,T., van Zwet,E.W., Dirks,R.W., Tanke,H.J., and van der Maarel,S.M. (2011). Reversible aggregation of PABPN1 pre-inclusion structures. Nucleus. *2*, 208-218.

Robinson,D.O., Wills,A.J., Hammans,S.R., Read,S.P., and Sillibourne,J. (2006). Oculopharyngeal muscular dystrophy: a point mutation which mimics the effect of the PABPN1 gene triplet repeat expansion mutation. J. Med. Genet. *43*, e23.

Sarkar,A.K., Biswas,S.K., Ghosh,A.K., Mitra,P., Ghosh,S.K., and Mathew,J. (1995). Oculopharyngeal muscular dystrophy. Indian J. Pediatr. *62*, 496-498.

Scacheri,P.C., Garcia,C., Hebert,R., and Hoffman,E.P. (1999). Unique PABP2 mutations in "Cajuns" suggest multiple founders of oculopharyngeal muscular dystrophy in populations with French ancestry. Am. J. Med. Genet. *86*, 477-481.

Schadt,E.E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature *461*, 218-223.

Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D., and Friedman,N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. *34*, 166-176.

Selbach,M., Schwanhausser,B., Thierfelder,N., Fang,Z., Khanin,R., and Rajewsky,N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63.

Semmler,A., Kress,W., Vielhaber,S., Schroder,R., and Kornblum,C. (2007). Variability of the recessive oculopharyngeal muscular dystrophy phenotype. Muscle Nerve *35*, 681-684.

Sharp,P.A. (2009). The centrality of RNA. Cell *136*, 577-580.

Sigal,A., Milo,R., Cohen,A., Geva-Zatorsky,N., Klein,Y., Liron,Y., Rosenfeld,N., Danon,T., Perzov,N., and Alon,U. (2006). Variability and memory of protein levels in human cells. Nature *444*, 643-646.

Soller,M. (2006). Pre-messenger RNA processing and its regulation: a genomic perspective. Cell Mol. Life Sci. *63*, 796-819.

Steele,E., Tucker,A., 't Hoen,P.A., and Schuemie,M.J. (2009). Literature-based priors for gene regulatory networks. Bioinformatics. *25*, 1768-1774.

Stuart,J.M., Segal,E., Koller,D., and Kim,S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249-255.

Tenenbaum,J.B., Kemp,C., Griffiths,T.L., and Goodman,N.D. (2011). How to grow a mind: statistics, structure, and abstraction. Science *331*, 1279-1285.

To,T.L. and Maheshri,N. (2010). Noise can induce bimodality in positive transcriptional feedback loops without bistability. Science *327*, 1142-1145.

Tome,F.M. and Fardeau,M. (1980). Nuclear inclusions in oculopharyngeal dystrophy. Acta Neuropathol. *49*, 85-87.

Uyama,E., Nohira,O., Tome,F.M., Chateau,D., Tokunaga,M., Ando,M., Maki,M., Okabe,T., and Uchino,M. (1997). Oculopharyngeal muscular dystrophy in Japan. Neuromuscul. Disord. *7 Suppl 1*, S41-S49.

Villagra,N.T., Bengoechea,R., Vaque,J.P., Llorca,J., Berciano,M.T., and Lafarga,M. (2008). Nuclear compartmentalization and dynamics of the poly(A)-binding protein nuclear 1 (PABPN1) inclusions in supraoptic neurons under physiological and osmotic stress conditions. Mol. Cell Neurosci. *37*, 622-633.

Wahl,M.C., Will,C.L., and Luhrmann,R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701-718.

Wahle,E. (1991). A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. Cell *66*, 759-768.

Wahle,E. (1995). Poly(A) tail length control is caused by termination of processive synthesis. J. Biol. Chem. *270*, 2800-2808.

Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P., and Burge,C.B. (2008a). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang,X., Arai,S., Song,X., Reichart,D., Du,K., Pascual,G., Tempst,P., Rosenfeld,M.G., Glass,C.K., and Kurokawa,R. (2008b). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature *454*, 126-130.

Wulff,B.E., Sakurai,M., and Nishikura,K. (2011). Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. Nat Rev Genet. *12*, 81-85.

MAKING SENSE OUT OF
TRANSCRIPTOMES

degree of similarity in diverse systems

PART ONE

Molecular and phenotypic characterization of a mouse model of oculopharyngeal muscular dystrophy reveals severe muscular atrophy restricted to fast glycolytic fibres

Capucine Trollet[1,2,3,4], Seyed Yahya Anvar[5], Andrea Venema[5], Iain P. Hargreaves[6], Keith Foster[1], Alban Vignaud[2,3,4], Arnaud Ferry[2,3,4], Elisa Negroni[2,3,4], Christophe Hourde[2,3,4], Martin A. Baraibar[7], Peter A.C. 't Hoen[5], Janet E. Davies[8], David C. Rubinsztein[8], Simon J. Heales[6], Vincent Mouly[2,3,4], Silvère M. van der Maarel[5], Gillian Bulter-Browne[2,3,4], Vered Raz[5] and George Dickson[1,*]

Oculopharyngeal muscular dystrophy (OPMD) is an adult-onset disorder characterized by ptosis, dysphagia and proximal limb weakness. Autosomal-dominant OPMD is caused by a short $(GCG)_{8-13}$ expansions within the first exon of the poly(A)-binding protein nuclear 1 gene (PABPN1), leading to an expanded polyalanine tract in the mutated protein. Expanded PABPN1 forms insoluble aggregates in the nuclei of skeletal muscle fibres. In order to gain insight into the different physiological processes affected in OPMD muscles, we have used a transgenic mouse model of OPMD (A17.1) and performed transcriptomic studies combined with a detailed phenotypic characterization of this model at three time points. The transcriptomic analysis revealed a massive gene deregulation in the A17.1 mice, among which we identified a significant deregulation of pathways associated with muscle atrophy. Using a mathematical model for progression, we have identified that one-third of the progressive genes were also associated with muscle atrophy. Functional and histological analysis of the skeletal muscle of this mouse model confirmed a severe and progressive muscular atrophy associated with a reduction in muscle strength. Moreover, muscle atrophy in the A17.1 mice was restricted to fast glycolytic fibres, containing a large number of intranuclear inclusions (INIs). The soleus muscle and, in particular, oxidative fibres were spared, even though they contained INIs albeit to a lesser degree. These results demonstrate a fibre-type specificity of muscle atrophy in this OPMD model. This study improves our understanding of the biological pathways modified in OPMD to identify potential bio-markers and new therapeutic targets.

---

1 Royal Holloway, University of London, Egham, UK. **2** UPMC Université Paris 6, UM76, France. **3** INSERM U974 and **4** CNRS UMR 7215, Institut de Myologie, Paris, France. **5** Center for Human and Clinical Genetics, Leiden University Medical Center, the Netherlands. **6** Neurometabolic Unit, National Hospital and Department of Molecular Neuroscience (Division of Neurochemistry), Institute of Neurology, Queen Square, London, UK. **7** Laboratoire de Biologie et Biochimie Cellulaire du Vieillissement, UR4, Université Pierre et Marie Curie, Paris 6, Paris, France. **8** Department of Medical Genetics, University of Cambridge, Cambridge Institute for Medical Research, Addenbrooke's Hospital, Cambridge, UK.

* To whom correspondence should be addressed at: g.dickson@rhul.ac.uk

Human Molecular Genetics, 2010, Vol. 19, No. 11, pp. 2191-2207

INTRODUCTION

Oculopharyngeal muscular dystrophy (OPMD) is a late-onset autosomal dominant genetic disease, characterized by progressive eyelid drooping, swallowing difficulty and proximal limb weakness in the late stages of the disease. The poly(A)-binding protein nuclear 1 (PABPN1) gene is mutated in OPMD patients and contains an expanded GCG trinucleotide repeat within exon 1 (Brais et al., 1998). This trinucleotide expansion is translated into a polyalanine tract at the N-terminus of the PABPN1 protein; in OPMD patients, this tract contains 12-17 alanine repeats instead of 10 repeats. PABPN1 with an expanded polyalanine tract forms nuclear aggregates (Tome et al., 1997). Although PABPN1 is ubiquitously expressed, the clinical and pathological phenotypes are restricted to skeletal muscles in OPMD patients, especially the pharyngeal and cricopharyngeal muscles (dysphagia), and the levator palpebrae superioris muscle (ptosis of the eyelid) (Perie et al., 2006).

PABPN1 is a protein localized in nuclear speckles, which binds with high affinity to poly(A) tails of mRNAs. PABPN1 promotes the interaction between the poly(A) polymerase and the cleavage and polyadenylation specificity factor, and controls the length of the poly(A) tail during polyadenylation of mRNA (Kuhn et al., 2009; Wahle, 1991; Wahle, 1995; Lemieux and Bachand, 2009). PABPN1 also contributes to the export of mRNA from the nucleus to the cytoplasm (Calado et al., 2000a; Apponi et al., 2010). The major pathological hallmark of OPMD in intranuclear inclusions (INIs) characterized by tubular filaments (Tome et al., 1997). It has previously been demonstrated that these INIs contain a large number of nuclear factors such as ubiquitin, subunits of the proteasome (Calado et al., 2000b), molecular chaperones HSP70 and HSP40 (Abu-Baker et al., 2003; Tavanez et al., 2009), poly(A) RNA (Calado et al., 2000b), protein involved in mRNA processing and transport CUGBP1, SFRS3, FKBP1A, hnRNP A1 and A/B and poly(A) polymerase (Corbeil-Girard et al., 2005; Fan et al., 2003; Tavanez et al., 2005). The exact role of PABPN1 aggregates in OPMD is still under debate. At present, it is still not clear whether the INIs observed in OPMD skeletal muscles have a pathological or a protective function by acting as a cellular defence mechanism against abnormal proteins. Several studies have suggested a pathological function of INIs: (i) the INIs could play a major role by sequestering essential cellular components such as specific mRNAs (Calado et al., 2000b) splicing or transcription factors (Corbeil-Girard et al., 2005; Fan et al., 2003), (ii) the frequency of INIs in nuclei of muscle fibres is correlated with the severity of the disease, with a frequency of 2-5% for heterozygous and 10% for homozygous patients (Blumen et al., 1999) and (iii) the reduction of the INIs in a mouse model by doxycycline or trehalose (Davies et al., 2006; Davies et al., 2005) or using intrabodies in a drosophila model (Chartier et al., 2009) improves muscle function. However, several studies have also suggested that the INIs might just be the result of a cellular defence mechanism and not the direct cause of the disease: (i) INIs are found both in affected and less-affected skeletal muscles, (ii) Tavanez et al. (2009) has recently proposed that the expansion alters the protein conformation and changes the binding properties of interacting proteins independently of the formation of INIs, (iii) the polyalanine domain of PABPN1 is not essential for aggregate formation (Tavanez et al., 2005; Chartier et al., 2006; Klein et al., 2008) and (iv) it has been suggested that the soluble form of the mutated PABPN1 is itself pathogenic, whereas the INIs would be a form of cellular protection (Catoire et al., 2008; Messaed et al., 2007).

In order to study the pathological mechanisms underlying OPMD, several in vitro models have been developed expressing an expanded PABPN1 transgene: transiently transfected COS-7 and HeLa cells (Abu-Baker et al., 2003; Messaed et al., 2007; Bao et al., 2002), adenovirus-infected A549tTA cells (Corbeil-Girard et al., 2005) or stably transfected C2 cells (Kim et al., 2001). In

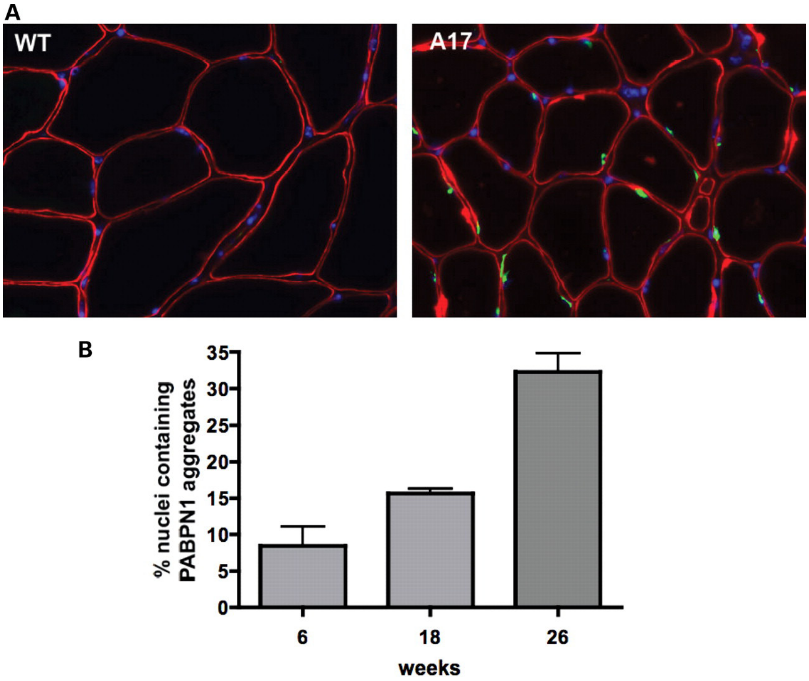


Figure 1 - A) The KCl-insoluble nuclear aggregates containing expPABPN1 (green) were detected by immunostaining on skeletal muscle cryosections from A17.1 mice. The sections of WT mice did not show any KCl-insoluble aggregates. (red, dystrophin; blue, nuclei; green, PABPN1; magnification ×400.) **B)** The percentage of nuclei containing PABPN1 aggregates was determined on skeletal muscle (TA) cryosections from 6 (T1), 18 (T2) or 26 (T3) weeks old A17 mice (n = 3 per time point with 250–350 fibres counted per muscle; the percentage of aggregates in T1 and T2 is significantly lower when compared with T3: T1 versus T2 ** P <0.01, T2 versus T3 *** P <0.001).

parallel, different animal models have also been generated: a drosophila model expressing PABPN1 with a polyalanine extension of different lengths, resulting in a muscular dystrophy with abnormal wing posture (Chartier et al., 2006), a nematode model expressing different lengths of expanded PABPN1 and showing muscle cell degeneration and abnormal mobility (Catoire et al., 2008) and several mouse models expressing either ubiquitously (Dion et al., 2005; Hino et al., 2004) or muscle specifically (Davies et al., 2005) expanded PABPN1 leading to the formation of INIs (Davies et al., 2005; Hino et al., 2004; Uyama et al., 2005). In the mouse model developed by Davies et al., a mutated version of PABPN1 with 17 alanines (expPABPN1) is expressed under the control of the human skeletal actin (HSA1) promoter, restricting the transgene expression to the striated muscle. Mice expressing the expPABPN1 transgene (A17.1) show a progressive muscle weakness and a progressive accumulation of INIs (Davies et al., 2005).

The aim of the present study was to gain insights into the different physiological pathways affected in OPMD muscles by performing both a general transcriptomic analysis and a detailed phenotypic characterization of the skeletal muscle of A17.1 mice compared with wild-type (WT) mice at different time points. We have observed that the muscle-restricted expression of the expPABPN1 transgene induces considerable gene expression deregulation among which genes associated with muscle atrophy were particularly affected. Functional and histological analysis of the skeletal muscle of this mouse further confirmed a severe muscular atrophy associated with

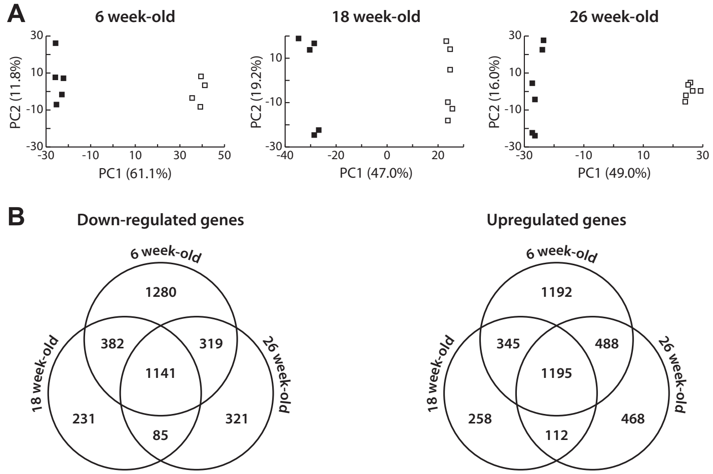


Figure 2 - Transcriptomic study in quadriceps muscles of A17.1 and WT mice at 6, 18 and 26 weeks. A) PCA plots for each time point data sets. A17.1 and WT samples are represented with black and white dots, respectively. **B)** Venn diagram of the deregulated genes using the unbiased cut-off P-value of 0.05, showing the number of A17.1 deregulated genes in each time point and the overlapping genes between two or three time points.

a reduction in muscle strength. Interestingly we showed that this muscular atrophy is restricted to fast glycolytic fibres, containing a large number of INIs, while oxidative fibres are spared, and contain less INIs. These results suggest a fibre-type specificity of muscle atrophy in this OPMD model, together with a less specific presence of INIs.

RESULTS

Gene expression profiling in muscle from mice expressing expPABPN1

To gain insight into molecular mechanisms involved in OPMD, we performed a transcriptomic analysis on skeletal muscle from A17.1 mice expressing an expanded form of PABPN1 with 17 alanines (expPABPN1). Davies et al. (Davies et al., 2005; Davies et al., 2008) previously described that these A17.1 mice show progressive formation of aggregates and progressive muscle weakness from approximately 18 weeks of age, whereas A10.1 mice expressing WT PABPN1 were indistinguishable from WT mice (Davies et al., 2008). By immunohistochemistry (**Figure 1A**), we confirmed that, in A17.1 mice, the number of nuclei containing PABPN1 increases with age. At 6 weeks (T1), 8% of the nuclei contained aggregates, and this number progressively increased to 15% at 18 weeks (T2) and 30% at 26 weeks (T3) (**Figure 1B**). Thus aggregation of expPABPN1 starts at a very early age, suggesting that potentially earlier muscle dysfunction may occur prior to the onset of muscle weakness symptoms observed from 18 weeks of age (Davies et al., 2005; Davies et al., 2008).

In order to identify the biological pathways that are initially deregulated, we carried out transcriptomic analyses on the skeletal muscle from 6-week-old mice (T1), when there are no obvious signs of muscle weakness, as well as from 18 (T2) and 26 weeks (T3) when the A17.1 mice are showing progressive muscle weakness. RNA expression arrays were generated from WT and A17.1 RNA isolated from quadriceps muscles, which were hybridized to Illumina Bead array v.1 containing 46632 unique probe identifiers. After normalization, the quality of the microarray hybridization was evaluated with the principal component analysis (PCA) (Chatterjee and Price, 1991; Pearson, 1901). For all three time points (T1-T3), PCA plots showed that mice with the same genotype (WT or A17.1) cluster together indicating that most variations in the arrays could be attributed to the genotype (**Figure 2A; PC1**). A weaker association was found with the second component representing technical variations. The Clustergrams representing hierarchical clustering for each time point (Supplementary Material, **Figure S1**) further demon-

Table 1 - Most significant A17.1 deregulated biological processes GO terms. Sorting is according to P-value.

ID	GO term	P-value	Genes	Deregulated genes
GO:0051169	Nuclear transport	1.35E-08	94	46 (49%)
GO:0009056	Catabolic process	1.41E-08	872	361 (41%)
GO:0015031	Protein transport	1.60E-08	591	250 (42%)
GO:0045859	Regulation of protein kinase activity	1.68E-08	132	50 (38%)
GO:0006796	Phosphate metabolic process	1.69E-08	804	313 (39%)
GO:0006950	Response to stress	1.77E-08	955	289 (30%)
GO:0006457	Protein folding	1.87E-08	107	47 (44%)
GO:0006397	mRNA processing	1.90E-08	219	111 (51%)
GO:0007049	Cell cycle	1.93E-08	615	197 (32%)
GO:0050790	Regulation of catalytic activity	1.96E-08	331	114 (34%)
GO:0006915	Apoptosis	2.03E-08	647	225 (35%)
GO:0051276	Chromosome organization and biogenesis	2.33E-08	319	135 (42%)
GO:0007517	Muscle development	2.49E-08	179	73 (41%)
GO:0009628	Response to abiotic stimulus	2.60E-08	185	60 (32%)
GO:0007005	Mitochondrion organization	2.68E-08	60	24 (40%)
GO:0006461	Protein complex assembly	2.89E-08	164	66 (40%)
GO:0010608	Posttranscriptional regulation of gene expression	2.27E-07	95	45 (47%)
GO:0006511	Ubiquitin-dependent protein catabolic process	2.27E-07	451	215 (48%)
GO:0016567	Protein ubiquitination	1.88E-05	53	26 (49%)
GO:0006412	Translation	9.69E-03	273	139 (51%)
GO:0042692	Muscle cell differentiation	9.80E-03	75	31 (41%)
GO:0048666	Neuron development	1.31E-02	276	83 (30%)

strated that the differences in the distribution of gene expression intensities between muscle samples from WT and A17.1 mice were due to changes in the individual gene expression levels between groups rather than nonspecific variations between samples. These results indicate that the gene expression changes between A17.1 and WT mice can be classified based on their genotype.

Subsequently, A17.1-deregulated genes were defined with a cut-off P-value of 0.05 and false discovery rate (FDR) corrected. The majority of up- or down-regulated genes were found in the 6-week-old mice (3220 and 3122, respectively). The total amount of either up- or down- regulated genes was gradually reduced at T2 (1910 and 1839, respectively) and T3 (2263 and 1866, respectively) (**Figure 2B**). This observation indicates that overexpression of the expPABPN1

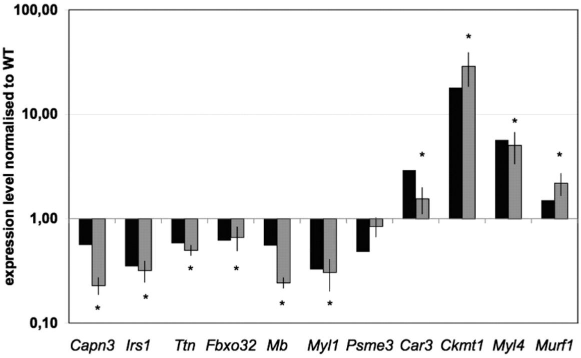


Figure 3 - Validation of A17.1 deregulated expression level of selected genes in skeletal muscle of A17 mice. Histograms indicate the expression levels normalized to that measured in the WT mice. Values measured by quantitative RT–PCR (greys bars) or microarray (black bars) are means ± standard deviations for n = 5–6 mice per group (* P <0.05).

gene leads to considerable changes in the expression of a large number of genes. More importantly, most of the A17.1 deregulated genes overlapped between two or three time points (T1 61%; T2 86.9%; T3 80.9%, **Figure 2B**), with a similar ratio between the up- or down-deregulated genes at all time-points, indicating that overexpression of expPABPN1 does not lead to preferential transcriptional up- or down-regulation in the A17.1 mouse.

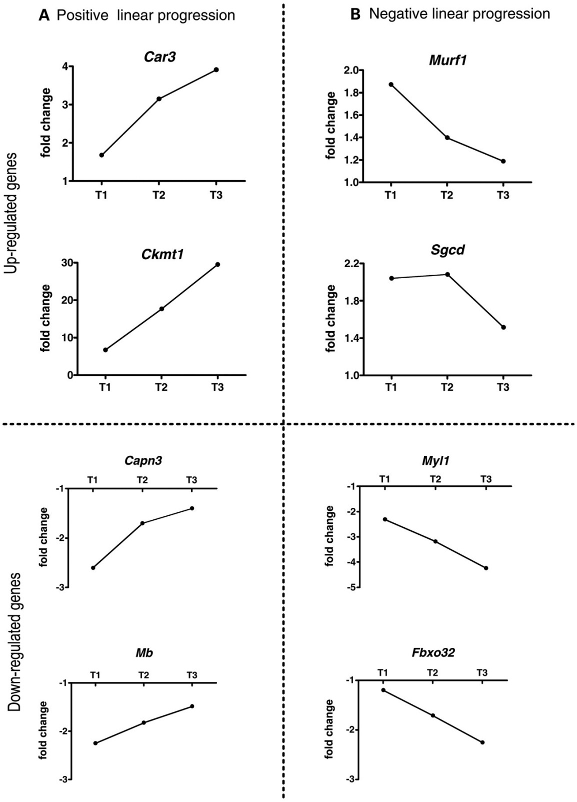


Figure 4 - Validation of the mathematical model for progression analysis. Expression plots of individual selected genes showed linear progression. The fold change is calculated from the microarray analysis. Graphs are sub-grouped according to up- or down-regulated genes and positive or negative linear regression.

Since a massive gene deregulation was found in the A17.1 mice, rather than taking a gene-by-gene analysis approach, we searched for the biological pathways that were significantly affected in this OPMD mouse model. To determine the gene ontology (GO) categories that were significantly associated with the expPABPN1 overexpression genotype, we used the global test analysis (Goeman et al., 2004). Significant GO categories were selected with the adjusted P-value of <0.05 corrected with FDR. Next, the significance of each GO term was evaluated using an enrichment analysis, which calculates the significance of each cluster based on the proportion of differentially expressed genes that contributes to the respective cluster. A list of biological GO categories that are significantly deregulated in the A17.1 mice was created using DAVID (Dennis, Jr. et al., 2003; Huang et al., 2009), revealing a broad range of deregulated biological processes in the A17.1 mice (**Table 1** and Supplementary Material **Table S1**). We identified transcriptional deregulation of genes involved in mRNA processing (GO:0006397), cell cycle (GO:0007049), the ubiquitin–proteasome pathway (GO:0006511 and GO:0016567), protein transport (GO:0015031) and the mitochondria (GO:0007005), corroborating a previous transcriptome analysis in an OPMD cell model (Corbeil-Girard et al., 2005). We also found a significant deregulation of apoptosis (GO:0006915), confirming the cell death previously described in this mouse model (Davies et al., 2005; Davies et al., 2008) and in a cellular model (Marie-Josee et al., 2006). Importantly, we found a significant deregulation of GO categories that affect muscle biology.

Since the A17.1-deregulated GO categories are biologically very broad and since OPMD affects muscle cells, we next used the literature to map significant biological concepts that would be muscle related. We assumed that the subgroup of overlapping deregulated genes common to the three time points is strongly associated with the disease aetiology, and therefore selected this subgroup for a literature-aided mapping of biological concepts using Anni 2.0 (Jelier et al., 2008). Out of the

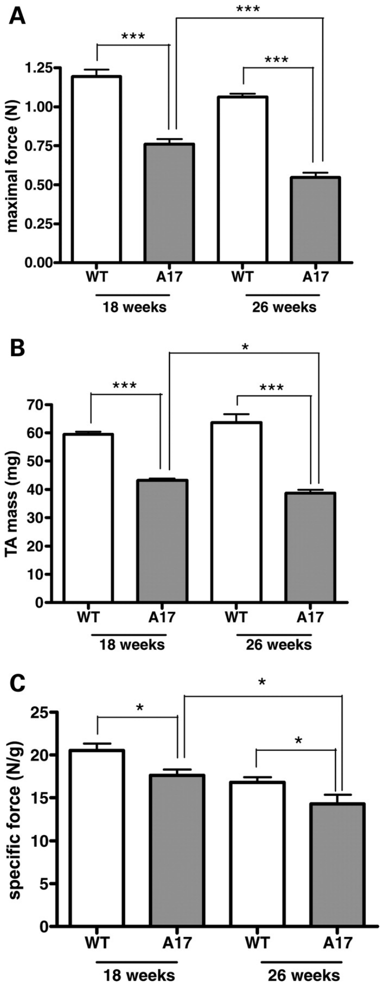


Figure 5 - Measurements of the weight and functional performance of skeletal muscle in WT and A17.1 mice at 18 and 26 weeks of age (n = 6 per group). A) The maximal force produced by the TA muscle was determined in WT and A17.1 mice (*** $P < 0.001$). **B)** The mass of the TA muscle was measured in A17.1 and WT mice (* $P < 0.05$; *** $P < 0.001$). **C)** The specific force (N/g) for the TA muscles of A17.1 and WT mice was calculated by dividing the maximal absolute force by the muscle mass (* $P < 0.05$).

2336 overlapping deregulated genes, only 1679 genes were recognized by Anni 2.0 (Supplementary Material **Table S2**). Among these, 481 deregulated genes (28.5%) were found to be highly associated with the terms 'muscle atrophy' or 'skeletal muscle atrophy' (Supplementary Material **Table S2**), suggesting that muscle atrophy may already be triggered in the A17.1 mouse at 6 weeks.

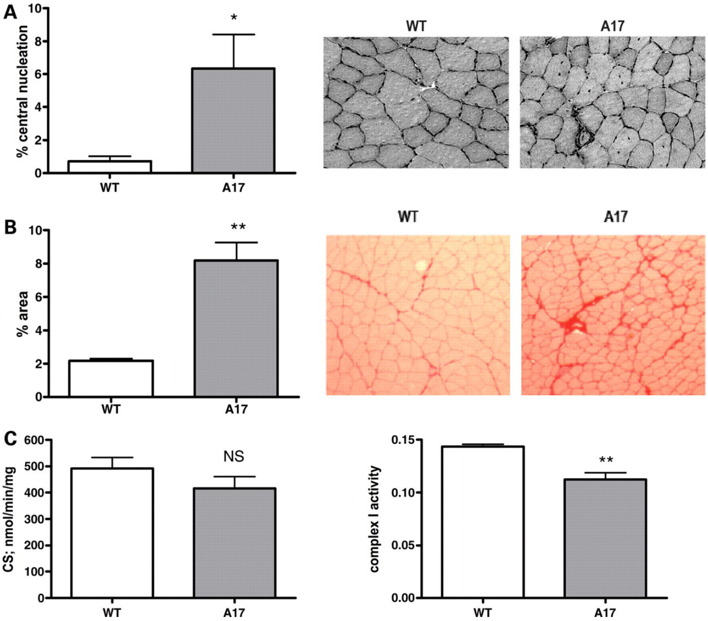
To validate the transcriptome analysis by quantitative PCR, we selected 10 genes from the muscular atrophy association list using both >1.3-fold change and high P-values criteria. RNA isolated from quadriceps of 6-week-old WT or A17.1 mice were used for the validation study. For each gene, the expression level was compared between the microarray and the quantitative PCR (**Figure 3**). After normalization to the WT control, a similar change in expression level was observed for each gene analysed, demonstrating that our microarray analysis is valid.

As muscle weakness in the A17.1 mice is progressive (Davies et al., 2005), we applied mathematical modelling for progressiveness on the A17.1-deregulated genes. A linear regression model was generated using the Limma model in R (Smyth, 2004) and was applied to all of the genes in the array. A total of 410 genes were identified as candidates for this progression. Subsequently, these 410 genes were applied in Anni 2.0 to find an association with the terms 'muscle atrophy' and 'skeletal muscle atrophy'. Of the 410 candidate genes, only 168 genes were available for Anni analysis. Among these 163 genes, 63 (38.6%) were highly associated with muscle atrophy in the biomedical literature (Supplementary Material **Table S2**). This analysis strongly suggests that the deregulation of muscle mass is progressive in the A17.1 mice. To confirm this analysis, eight genes were selected using the fold change criteria and their expression profiles over time were plotted. The progression plots of fold change showed a linear positive or negative progression for all selected up- or down-regulated genes, therefore validating the mathematical modelling (**Figure 4**).

Muscle atrophy in A17.1 mice

Since the transcriptomic study indicates muscle atrophy in the A17.1 mice, we performed a detailed analysis of the skeletal muscles of the A17.1 mice over time. Using the grip test, it was previously shown that the A17.1 mice develop a progressive muscle weakness with a significant decrease in strength compared with WT mice from 18 weeks of age (Davies et al., 2005; Davies et al., 2008), whereas A10.1 mice expressing WT PABPN1 were indistinguishable from WT mice (Davies et al., 2008).

Figure 6 - A) Centrally nucleated fibres were determined on transversal sections of WT (n = 3) and A17.1 (n = 4) TA muscles of 26 weeks old mice following hematoxylin/eosin staining. For each section, more than 800 fibres were counted from four random areas. The results represent the percentage of centrally nucleated fibres (* P <0.05). **B)** Sirius Red staining of transversal sections of WT (n = 3) and A17.1 (n = 4) TA muscles of 26 weeks old mice (** P <0.01). **C)** Citrate synthase (CS) activity and mitochondrial complex I activity measurement on WT and A17.1 muscle (n = 6 per group) from 26-week-old mice. The activity of complex I is expressed in nmol/min/ml and then normalized relative to citrate synthase as an indicator of mitochondrial content (** P <0.01).



In order to further analyse the consequences of the expPABPN1 expression on the physiological function of skeletal muscle, we measured the contractile properties of the tibialis anterior (TA) skeletal muscle of A17.1 mice at 18 and 26 weeks of age when compared with age-matched WT littermates. The maximal absolute force of the TA of A17.1 transgenic mice was significantly reduced by 36% at 18 weeks and 48% at 26 weeks when compared with WT mice (**Figure 5A**). The mass of the TA muscle of A17.1 mice was progressively reduced by 27% at 18 weeks and 39% at 26 weeks when compared with WT mice (**Figure 5B**). This progressive reduction in muscle mass was observed from 6 weeks of age (20% reduction at 6 weeks, data not shown) and was also observed in other skeletal muscles such as the quadriceps and the gastrocnemius (data not shown). This effect is specifically due to the overexpression of expPABPN1 since A10.1 mice expressing WT PABPN1 at higher level than A17.1 mice did not show a similar reduction in TA muscle mass (data not shown). We next calculated the specific force of the TA muscles of A17.1 and WT mice by normalizing the maximal (absolute) force to the muscle mass. This measure showed that the specific force of the TA of A17.1 mice was significantly reduced by 14% at 18 weeks and 20% at 26 weeks when compared with WT mice (**Figure 5C**). Both decreases in muscle mass and in specific force participate to the decrease in absolute maximal force.

This reduced specific force demonstrates that there is both a qualitative change as well as an additional pathological process occurring in the skeletal muscle. Whereas immunostaining on muscle sections did not reveal any obvious modifications of the dystrophin-associated glycoprotein complex (data not shown), an haematoxylin-eosin (H&E) staining revealed an increased number of centrally nucleated fibres in A17.1 when compared with WT mice (**Figure 6A**). In addition, a Sirius red staining revealed a more pronounced endomysial fibrosis in A17.1 mice when compared with WT mice (**Figure 6B**), which could potentially explain the reduced specific force. This muscle weakness could also result from a modified mitochondrial function, as this pathway was shown to be deregulated in the transcriptomic data (**Table 1**). Mitochondrial ATP is generated via oxidative phosphorylation through the combined action of five enzyme complexes. Ci-

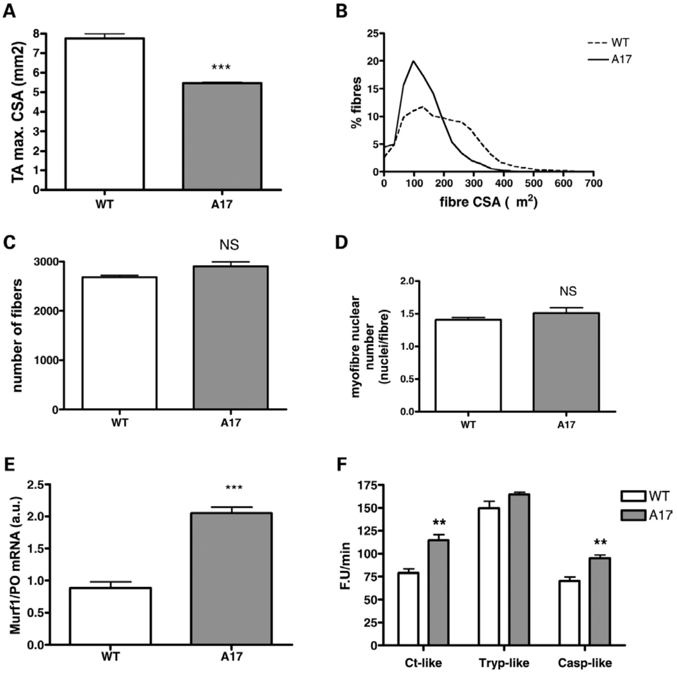
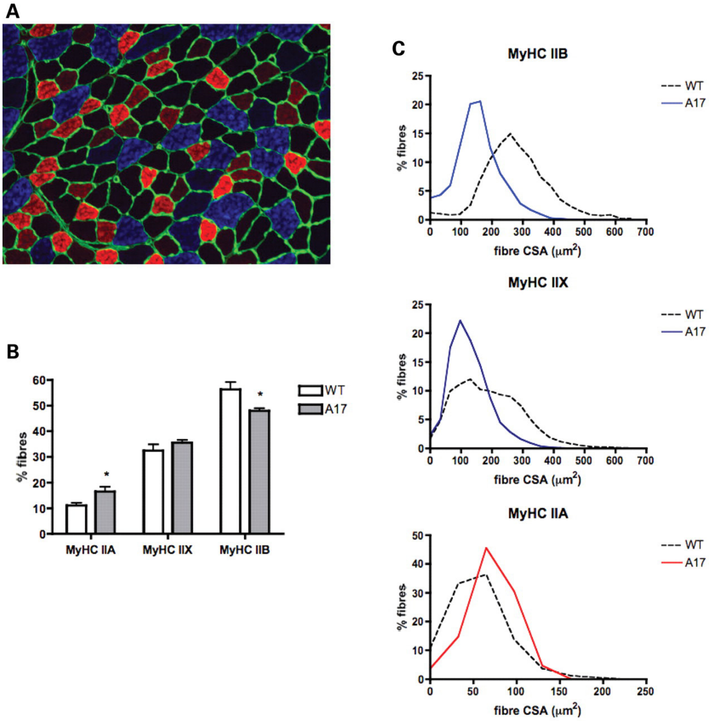


Figure 7 - Evaluation of the skeletal muscle atrophy at 26 weeks of age. A) The maximal cross section area (CSA) of the TA muscle was measured for the TA of WT and A17.1 mice (n = 6 per group), ∗∗∗ P <0.001. **B)** The frequency of the cross-sectional area (CSA) of the muscle fibres was determined in the TA muscle from WT and A17.1 mice. The plotted lines represent the mean of three different muscles for each group (x2 analysis performed on data sets, P <0.001). **C)** The total number of fibres per muscle was determined in the TA of WT and A17.1 mice and did not show any difference (n = 3 per group). NS represents non-significant. **D)** The number of nuclei per fibre on TA muscle section was similar in both A17.1 and WT mice (n = 4 per group with around 250–350 fibres counted per muscle). **E)** MuRF1 mRNA expression in TA muscle of 26-week-old mice. Values measured by quantitative RT–PCR are means ± standard deviations for n = 5–6 mice per group, ∗ P <0.05. **F)** Proteasome activity in TA muscle of 26-week-old mice. Ct-like, chimotrypsin-like; Tryp-like, Trypsin-like; Casp-like, Caspase-like. The results are expressed in F.U/min, ∗∗ P <0.01; n = 4 for WT and n = 6 for A17.1.

trate synthase levels were similar in WT and A17.1 TA muscle, suggesting that the total amount of mitochondria is preserved; however, assessment of mitochondrial respiratory chain enzyme activity showed a decreased activity in complex I (NADH: ubiquinone reductase) in A17.1 when compared with WT mice (**Figure 6C**). In contrast, complex II–III (succinate:cytochrome c reductase) and complex IV (cytochrome c oxidase) activities were not decreased when compared with control levels (data not shown). These data suggest a mitochondrial dysfunction, which could result in muscle contractile defects and therefore also participate to the decrease in specific force.

The reduction in muscle mass and force together with the transcriptomic data suggests that the expression of expPABPN1 triggers muscular atrophy. To further confirm this hypothesis, we performed a detailed histological analysis of the TA muscle of A17.1 mice when compared with WT mice at 26 weeks of age. We observed a 30% reduction in the maximal cross-sectional area (CSA) of the TA in A17.1 transgenic mice when compared with their age-matched WT littermates (**Figure 7A**). On muscle sections that generated the maximal CSA, we subsequently analysed individual fibre CSA using an anti-laminin antibody to delimit the muscle fibres. When the frequency distribution of the fibres was plotted according to their CSA (**Figure 7B**), a shift was observed from the large towards the small size of muscle fibres in the A17.1 mice. The CSA was reduced by ~30% (189 mm2 for WT mice and 132 mm2 for A17.1 mice), whereas there was no change in the total number of fibres between WT and A17.1 mice (**Figure 7C**). Interestingly, the reduction in muscle size was not associated with a decrease in myonuclear number (**Figure 7D**). Overall, these results confirm muscular atrophy, defined as a decrease in cell size by loss of organelles, cytoplasm and proteins (Sandri, 2008). This reduction in muscle mass is due to an

Figure 8 - The myosin heavy chain (MyHC) muscle fibres subtypes were determined by immunostaining. A) Immunostaining of laminin (green), MyHC-IIA (red), MyHC-IIB (blue) on a TA muscle cryosection. The distribution of muscle fibre subtypes (**B**) and the frequency of the cross-sectional area (**C**) of each muscle fibre subtype were determined in the whole of TA muscle from WT and A17.1 mice at 26 weeks. The data represented are the mean of three different muscles for each group. ∗ P <0.05; and X^2 analysis performed on data sets: MyHC-IIX P <0.001 and MyHC-IIB P <0.001.



---

improper balance between protein synthesis and degradation, inducing a loss of total protein content in muscle fibres (Nury et al., 2007). The ubiquitin–proteasome pathway is activated during muscle atrophy and is involved in the breakdown of major contractile proteins (Gilson et al., 2007; Vazeille et al., 2008). In particular, MuRF-1 and Atrogin-1, known as atrogenes, play a crucial role in the loss of muscle proteins and their expression is considered as specific atrophy markers (Sandri, 2008; Bodine et al., 2001). In the progression analysis (**Figure 4**) and by quantitative RT–PCR (**Figure 3**), we have shown that the atrogene MuRF-1 only was indeed up-regulated in A17.1 mice. We further confirmed that in the TA of 26-week-old A17.1 mice there was a persistence of this MuRF-1 mRNA up-regulation (**Figure 7E**), mainly mediated by a down-regulation of the active phosphorylated form of PKB/Akt and a translocation of Foxo3A transcription factor to the nucleus (Supplementary Material **Figure S2**). We also measured proteasome activities (chymotrypsine-, trypsin- and caspase-like) in the TA muscle of 26-week-old mice and observed a significant increase in the chymotrypsinand caspase-like activity in A17.1 mice, whereas the trypsinlike activity was not significantly increased (**Figure 7F**). Altogether, these data confirm muscular atrophy in the A17.1 mice.

In order to further evaluate if we could locally reproduce this atrophic phenotype in the skeletal muscle of an adult WT mice, we overexpressed the expanded PABPN1 transgene using an adeno-associated virus (rAAV2/8-CAGexpPABPN1, Supplementary Material **Figure S3A**) injected into the TA of WT mice at 8 weeks of age. Three months post-injection, we confirmed the overexpression of expPABPN1 and the presence of expPABPN1 INIs only in the injected muscle fibres (Supplementary Material **Figure S3B**). Similar to what we measured in A17.1 mice, we observed a reduced muscle mass and reduced maximal force of the injected TA of WT mice when compared with the contralateral un-injected leg, leading to a slight but not significant reduction

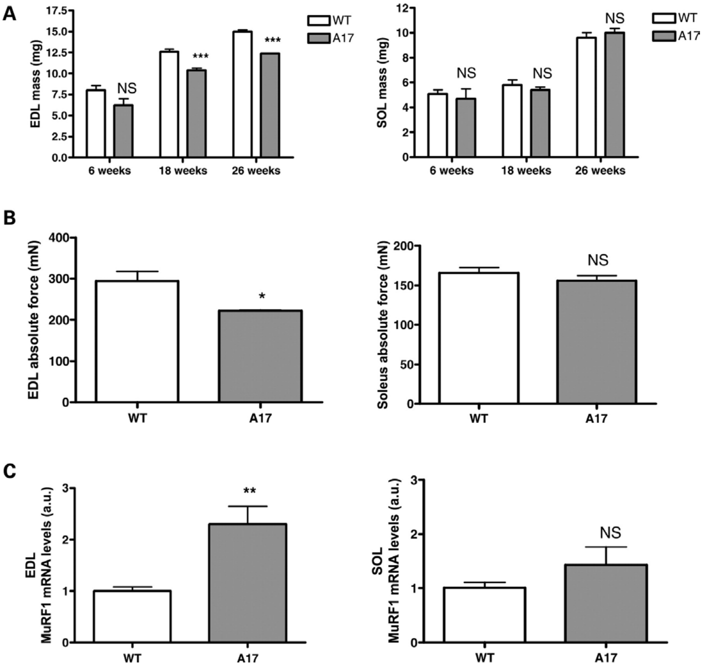


Figure 9 - The weight and functional performance of the soleus (SOL) and extensor digitorum longus (EDL) was evaluated in A17.1 and WT mice. A) The muscle mass of the SOL and EDL muscles of A17.1 and WT mice was measured at 6, 18 and 26 weeks of age (6 weeks n = 4 per group; 18 weeks n = 6 per group; 26 weeks n = 6 per group; *** P <0.001; NS is non-significant). **B)** The maximal force of SOL and EDL was evaluated for A17.1 and WT mice at 26 weeks (* P <0.05; NS is non-significant). **C)** MuRF1 mRNA expression in EDL and soleus muscle from 6-week-old A17.1 and WT mice. Values measured by quantitative RT–PCR are means ± standard deviations for n = 4–6 mice per group (** P <0.01; NS is non-significant).

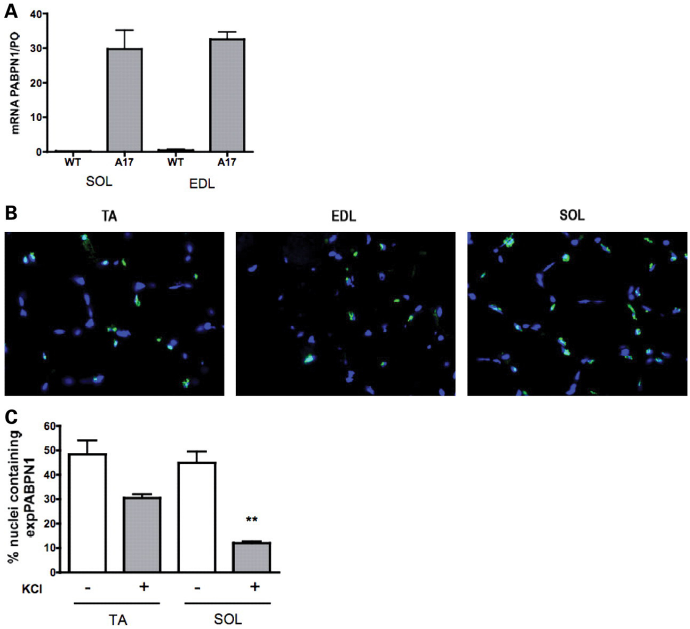
in the specific force (Supplementary Material **Figure S3C**). This result further confirms that the expression of expPABPN1 in mature muscle fibres induces an atrophic process.

Distinct phenotypes between oxidative and glycolytic fibre subtypes

Muscle is composed of distinct fibre types, which can be defined by the myosin heavy-chain isotypes (MyHC) they express: MyHC-I in the slow oxidative fibres, MyHC-IIA in the fast oxidative fibres, MyHC-IIX and MyHC-IIB in the fast glycolytic fibres (Bottinelli and Reggiani, 2000). We investigated whether the distribution and CSA of oxidative/glycolytic muscle fibre subtypes in the TA muscle was modified in A17.1 mice compared with WT mice. We therefore performed a co-immunostaining of the different myosin heavy chains together with a laminin staining to determine the CSA (**Figure 8A**). The distribution analysis revealed that the A17.1 muscles had more MyHC-IIA fibres (17 versus 9% in WT muscle) and fewer MyHC-IIB (48 versus 56% in WT muscle) (**Figure 8B**). Interestingly, this result is in accordance with the down-regulation of Myl1 mRNA (fast myosin light chain) observed by quantitative PCR (**Figure 3**). By plotting the frequency distribution of CSA myofibre subtypes, we observed a shift towards the small size for the specific MyHC-IIB and MyHC-IIX fibres, whereas surprisingly the MyHC-IIA fibres were unaffected (**Figure 8C**). This result suggests that the fast glycolytic fibres are specifically affected in the A17.1 mice.

In order to further confirm this selective muscle atrophy of the fast glycolytic fibres, we analysed two other muscles: the extensor digitorum longus (EDL) muscle considered as a ‘fast’ muscle type and composed of MyHC-IIA, -IIX and -IIB fibres like the TA, and the soleus (SOL) muscle, considered as a mixed muscle type and composed of MyHC-I and MyHC-IIA fibres. As shown in Figure 9A, the muscle mass of the EDL of A17.1 mice was reduced by 20% when compared with WT mice, whereas the muscle mass of the SOL was unchanged in A17.1 and WT mice at 6, 18 and 26 weeks. This difference between the SOL and the EDL further suggests that muscle atrophy

Figure 10 - expPABPN1 expression in TA, EDL and SOL muscles before and after KCl treatment. A) Quantitative RT–PCR on EDL and SOL muscle at 6 weeks of age (n = 3 for WT; n = 6 for A17 samples). **B)** Immunostaining of expPABPN1 in muscle cryosections (TA, EDL and SOL) without any KCl treatment (expPABPN1 in green and nuclei stained with Hoechst in blue). **C)** Amount of expPABPN1 positive nuclei before and after KCl treatment to remove any soluble protein (n = 4 per group with around 250–350 fibres counted per muscle, TA and SOL (** P <0.01 ANOVA test).



in A17.1 mice is restricted to fast glycolytic fibres. The maximal force measurement of these two muscle types revealed a decrease in force for the EDL of A17.1 mice, whereas for the SOL muscle, we did not observe any difference in the maximal force between WT and A17.1 mice (**Figure 9B**), confirming that the SOL muscle is spared. By quantitative PCR we further observed in the EDL muscle a 2-fold increase in MuRF-1 expression similar to previous results observed in quadriceps and TA muscle, whereas we did not observe any statistical difference for MuRF-1 expression in the soleus muscle (**Figure 9C**).

Since we found selective muscle atrophy of the EDL but not of the SOL muscle, we determined whether such a difference could be due to differential transgene expression levels. The transgene is under the control of the HSA promoter, which restricts the transgene expression to skeletal muscles, including the SOL as well as the EDL (Brennan and Hardeman, 1993; Miniou et al., 1999; Orengo et al., 2008). By quantitative RT–PCR, we confirmed that there were equal mRNA expression levels in both the SOL and EDL muscles (**Figure 10A**). Thus differential transgene expression cannot explain the selective muscle involvement. Therefore, we continued by performing a direct comparison by immunohistochemical staining of expPABPN1 expression in EDL and SOL muscle sections of 26-week-old A17.1 and WT mice. The PABPN1 immunostaining revealed a similar pattern of expression in EDL, SOL and TA with a high PABPN1 signal observed in around 45% of the nuclei in all muscle types (**Figure 10B**). Interestingly, when we performed a KCl treatment to remove soluble proteins, the amount of aggregates in the SOL was higher than in WT, but still two-fold lower than the levels observed in the TA of A17.1 animals (**Figure 10C**). Our results thus demonstrate that muscle atrophy in A17.1 mice is specific to fast glycolytic fibres and that these fibres contain a larger number of KCl-resistant INIs. In slow and fast oxidative fibres that do not show muscle atrophy, fewer INIs are observed.

DISCUSSION

The aim of the present study was to gain further insight into the biological pathways modified in OPMD muscles by a combination of transcriptomic and physiological studies. To generate a comprehensive picture of the deregulated pathways during disease progression in this mouse model, we have selected three time points for the transcriptomic analysis: 6 weeks as an early time point before onset of disease symptoms and 18 and 26 weeks when the mice show progressive muscle weakness (Davies et al., 2005; Davies et al., 2008). We observed a massive gene deregulation in A17.1 mice when compared with WT mice at all three time points. Among the GO terms revealed in this study, we identified several pathways deregulated such as mRNA processing, cell cycle, protein transport, mitochondria and apoptosis, which corroborate a previous gene-based transcriptome analysis of an in vitro OPMD cell model (Corbeil-Girard et al., 2005). We also found a deregulation of genes involved in muscle development and muscle cell differentiation, which could potentially emphasize defects in continuous remodelling of muscle, previously demonstrated in OPMD (Kim et al., 2001; Wirtschafter et al., 2004; Mouly et al., 2005). By mapping the biological concepts associated with this deregulation, we found that the muscle-restricted expression of expPABPN1 induced major and progressive deregulation of genes associated with muscle atrophy. Skeletal muscle atrophy is characterized by a decrease in muscle mass and consequently reduced contractile force of the muscle. Functional and histological analysis of the skeletal muscle of this mouse model confirmed severe muscular atrophy associated with a reduction in muscle strength. This atrophic phenotype was due specifically to the overexpression of the alanine expanded PABPN1 and not simply to overexpression of PABPN1 as we did not observe a severe muscle atrophy in the A10.1 mice expressing WT PABPN1. In accordance with this result, genes associated with atrophy such as MuRF-1 were not changed in the A10.1 mice (data not shown). The transcriptomic analysis showed homology with previous studies describing the transcriptional changes involved in muscle atrophy (Jagoe et al., 2002; Lecker et al., 2004; Sacheck et al., 2007; Calura et al., 2008), such as increased expression of atrogenes involved in protein degradation and decreased expression of genes involved in energy production. Two major pathways mediate protein degradation in skeletal muscle: the autophagic/lysosomal pathway and the ubiquitin-proteasomal pathway (UPP). In the A17.1 skeletal muscles, we confirmed at all time-points up-regulation of the muscle-specific ubiquitin ligase MuRF-1 gene expression. Since MuRF-1 is a known atrogene playing a crucial role in the loss of muscle proteins (Sandri, 2008; Bodine et al., 2001; Clarke et al., 2007; Cohen et al., 2009; Kedar et al., 2004), these data together with the increased proteasome activity in A17.1 muscles suggest an increased protein degradation rate in A17.1 mice related to muscle atrophy. These data also further support previous studies, which showed that the proteasome is thought to be the major degradation pathway for PABPN1 (Abu-Baker et al., 2003; Davies et al., 2006). Interestingly, MuRF-1 has also been described to be a potential energy homeostasis regulator for muscle (Hirner et al., 2008). Together with the deregulation of genes involved in protein degradation, we also observed a deregulation of genes involved in energy production—as described in other atrophic conditions (Jagoe et al., 2002; Lecker et al., 2004; Sacheck et al., 2007; Calura et al., 2008)—among which a significant cluster of genes related to mitochondrial organization. We observed a decreased mitochondrial respiratory chain complex I activity in skeletal muscle of the A17.1 mice. This suggests some impairment of oxidative phosphorylation that may contribute to the muscle dysfunction observed in this mouse model of OMPD. This is of particular interest since mitochondrial abnormalities have frequently been observed in OPMD patients (Muqit et al., 2008; Pauzner et al., 1991; Schroder et al., 1995). This decrease may solely be the result of a deregulation of genes encoding several subunits of complex I, as observed both in our transcriptomic data and in the previous transcriptomic analysis performed in an OPMD cell culture model (Corbeil-Girard et al., 2005). Respiratory chain enzymes are also susceptible to free radical-induced oxidative damage (Zhang et al., 1990), therefore an increased oxidative stress may also contribute to the decreased complex I activity, as suggested in the transcriptomic analysis (response to oxidative stress, GO:0006979). Toriumi et al. (2008) have recently shown that the polyalanine tract may induce mitochondrial dysfunction with the rupture of the mitochondrial membrane, release of cytochrome c and apoptosis (Toriumi et al., 2009). We demonstrated here that the reduction in muscle force was not just a consequence of muscle atrophy—as observed with the reduced specific force—so expPABPN1 expression clearly has a deleterious effect in force production potentially via mitochondrial dysfunction or oxidative stress, which will both need to be studied in more detail.

Interestingly, the detailed characterization of the skeletal muscle phenotype of these mice revealed a selective atrophy of the fast glycolytic fibres that contained the highest number of INIs, whereas the oxidative fibres containing less INIs were spared. This result suggests fibre-type specificity for both muscular atrophy and INIs formation in OPMD, indicating that depending on the muscle metabolic properties, the expression of expPABPN1 leads to different phenotypes. This raises two questions: why are there more INIs in fast glycolytic fibres? And why are oxidative fibres not affected even if these fibres contain INIs? The presence of INIs in both affected (EDL) and non-affected muscles (SOL) further emphasize the complex and poorly understood role of INIs in OPMD, which is still currently under debate. Whereas several studies have suggested a pathological function of INIs, several other studies have suggested that the INIs might just be the result of a cellular defence mechanism and not the direct cause of the disease. In this OPMD mouse model, we observed before KCl treatment similar amount of expPABPN1 expression in affected and unaffected muscles, which suggests that the soluble form of the protein in oxidative fibres is not toxic. We also observed that fast glycolytic fibres contained progressively larger numbers of INIs and were progressively atrophied, which could support the pathological function of aggregates. However, the presence of unaffected oxidative fibres containing INIs suggest that INIs are not the only factor involve in muscle atrophy. The difference in the amount of INIs will need to be more extensively studied to understand why more aggregates are found in fast glycolytic fibres when compared with slow oxidative fibres. There might be a fibre-type-specific mRNA/protein preventing (in oxidative fibres) or enhancing (in glycolytic fibres) the formation of INIs, or these two muscle fibre types may have a different protein degradation system. These two hypotheses need to be evaluated in the future. We also have to keep in mind that oxidative fibres seems to be more resistant to atrophy through a protective mechanism mediated by enhanced antioxidant gene expression (Sandri, 2008; Li et al., 2007; Yu et al., 2008), and therefore might be more resistant to the presence of expanded PABPN1. Another possible mechanism for this selective atrophy is based on the fact that nuclei in slow fibres contain a smaller myonuclear domain than fast fibres (Bruusgaard et al., 2006; Gundersen and Bruusgaard, 2008); so nuclear defects could potentially have fewer consequences and be less visible in slow fibres.

To summarize, we have shown that expression of expPABPN1 in muscle fibres leads to a massive gene deregulation with muscle atrophy as a major consequence. The muscle weakness we have observed results both from a reduction in muscle mass and a muscle dysfunction due to increased fibrosis, mitochondrial defects and possible oxidative stress. At the fibre-type level, we showed that only glycolytic fibres containing the largest number of INIs were affected, whereas oxidative fibres were spared and contained less INIs. In conclusion, expression of mutant PABPN1 in skeletal muscle of the A17.1 mouse recapitulates several pathological observations seen in OPMD patients: progressive muscle weakness, muscle atrophy, fibrosis, mitochondrial defects, affected and unaffected muscle containing INIs. These molecular and pathological changes will improve our understating of the disease progress in OPMD patients and should provide targets for future therapeutic strategies that may reverse some or all of these modified pathways essential for muscle homeostasis and normal function.

MATERIALS AND METHODS

Mice

A17.1 transgenic mice have previously been described (Davies et al., 2005). Male A17.1 mice and WT controls were generated by crossing the heterozygous carrier strain A17.1 obtained from Rubinsztein's group (Davies et al., 2005) with the FvB background mice. The mice were genotyped by PCR 3–4 weeks after birth. Wild type FvB and A17.1 mice were housed in minimal disease facilities (Royal Holloway, University of London) with food and water ad libitum.

RNA isolation and microarray processing

Total RNA was extracted from skeletal muscles using RNA Bee (Amsbio) according to the manufacturer's instructions. RNA integration number (RIN) was determined with RNA 6000 Nano (Agilent Technologies). RNA with RIN >7 were used for subsequent steps. RNA labelling was performed with the Illumina TotalPrep RNA Amplification kit (Ambion) according to the manufacturer's protocol, and subsequently was hybridized to Illumina Mouse v1.1 Bead arrays.

Data processing and analysis

Before data analysis, microarray measurements were normalized to remove systematic errors by balancing the fluorescence intensities using the quantile method (Smyth and Speed, 2003). Each time point has been normalized separately. Next, PCA plots were generated to assess the quality of the data (Chatterjee and Price, 1991; Pearson, 1901). This analysis showed that 47% of the variations within each data set were attributed to the genetic variation between the WT and the transgenic mice. Subsequently, statistical analysis was conducted using limma package in R (Smyth, 2004) to identify genes with significant differences in expression pattern between A17.1 and WT. Statistical analysis includes a cut-off P-value of 0.05 and FDR correction provided in the limma package in R. Probe annotation was made with the Illumina mouse whole-genome bead array version 1 annotation package.

GO analysis. The illuminaMousev1BeadID was used to describe the gene clustering arrangements based on the vocabulary of GO. These clusters have been used to conduct the significance of GO terms using global test (Goeman et al., 2004) by assigning a P-value to each cluster based on the assessment of how well group labels can be predicted for different samples (A17.1 versus WT) based on a regression model. The significance of these GO terms was validated using enrichment analysis. Enrichment analysis uses a hypergeometric test to calculate the significance of each cluster based on the number of differentially expressed genes it holds. In this study, we preferred global test for assessing the significance of GO terms over enrichment method due to an unrealistic assumption in which genes are treated as black and white (differentially or non-differentially expressed) for conducting the significance of each GO category whereas, in global test, gene expression profiles are being used to conduct such an analysis. Subsequently, DAVID functional annotation clustering tool (Dennis, Jr. et al., 2003; Huang et al., 2009) has been applied to remove redundancy and increase the specificity threshold for selected pathways, and finally, the list of deregulated genes was mapped to the concepts in biomedical literature using Anni 2.0 (Jelier et al., 2008). GO categories were selected based on the combination of the following criteria (1): GO categories with the adjusted P-value of <0.05; (2) clusters of GO categories generated by DAVID, which have P-values >0.05 will be discarded from the analysis; (3) GO categories that contain at least five genes and less than 1000; (4) from each cluster of GO categories, generated by DAVID, only two were selected for follow-up studies to reduce the redundancy. Subsequently, the 2336 genes that were differentially expressed throughout all three time points were mapped to biomedical concepts using Anni 2.0.

Definition of muscle atrophy-related genes. Muscle atrophy-related genes are defined as differentially expressed genes associated with the term 'muscle atrophy' in the biomedical literature, as determined with the literature analysis tool Anni 2.0 with the association score larger than 0.005.

Real-time RT–PCR analysis

Primers for validation were selected from the gene sequence that harbours the Illumina probe location using Primer 3 plus program. RNA was extracted using RNA Bee (Amsbio) and treated with RQ1 RNase-Free DNase (Promega). Subsequently, RNA was reverse transcribed using RevertAid H Minus M-MuLV First Strand kit (Fermentas) according to the manufacturer's instructions. An amount of 3.6 ng of cDNA was used for quantitative PCR using SYBR green mix buffer (BioRad) in a total of 15 ml reaction volume. PCR was carried out as follows: 4 min at 95°C followed by 40 cycles at 95°C for 10 s and 60°C for 45 s, the program ended in 1 min at 95°C and 1 min at 60°C. Specificity of the PCR product was checked by melting-curve analysis using the following program: 65°C increasing 0.5°C in 60 steps of 10s duration. Expression levels were calculated according to the DDCt method normalized to the mHPRT mRNA expression and to the average of the gene expression level in the WT mice. The statistical significance was determined with Student's t-test.

Measurement of muscle contractile properties

Contractile properties of TA muscle were evaluated by measuring the in situ isometric muscle contraction in response to nerve stimulation as described previously (Vignaud et al., 2007). Mice were anaesthetized using a pentobarbital solution (i.p. 60 mg/kg). The knee and foot were fixed with clamps and the distal tendons of the muscles were attached to an isometric transducer (Harvard Bioscience) using a silk ligature. The sciatic nerves were proximally crushed and distally stimulated by a bipolar silver electrode using supramaximal square-wave pulses of 0.1 ms duration. All data provided by the isometric transducer were recorded and analysed using PowerLab system (4SP, AD Instruments). All isometric measurements were made at an initial length L0 (length at which maximal tension was obtained during the twitch). Responses to tetanic stimulation (pulse frequency from 6.25, 12.5, 25, 50, 100 and 143 Hz) were successively recorded and the maximal force was determined. After contractile measurements, mice were sacrificed with an overdose of anaesthetic solution. Muscles were then weighed to calculate the specific maximal force, frozen in isopentane cooled in liquid nitrogen and stored at ~80°C.

The isometric contractile properties of soleus and extensor digitorum longus muscles were studied in vitro. Measurements were performed as described previously (Vignaud et al., 2008). The muscles were dissected free from adjacent connective tissue and soaked in an oxygenated Tyrode solution (95% O_2 and 5% CO_2) containing (mM): NaCl (118), $NaHCO_3$ (25), KCl (5), KH_2PO_4 (1), $CaCl_2$ (2.5), $MgSO_4$ (1), glucose (5), and maintained at a temperature of 20°C. Muscles were connected at one end to a force transducer. After equilibration (30 min), electrical stimulation was delivered through electrodes running parallel to the muscle. Isometric contractions were recorded at the length at which maximal isometric tetanic force was observed (L_0). Absolute maximal isometric force (mN) was measured (usual frequency of 125 Hz, train of stimulation of 1500 ms). Specific maximal force (mN/mm^2) was calculated by dividing the force by the estimated CSA of the muscle. Assuming that muscles have a cylindrical shape and a density of 1.06 mg mm^{-3}, muscle CSA corresponds to the wet weight of the muscle divided by its fibre length (L_f).

The fibre length to L_0 ratio of 0.70 (soleus) or 0.45 (EDL) was used to calculate L_f. Muscles were weighed and frozen in liquid nitrogen.

Muscle histology, immunohistochemistry and morphometric measurements

Recovered tissues were mounted in Cryo-M-Bed (Bright Instruments, Huntingdon, UK) and snap frozen in liquid nitrogen-cooled isopentane. Staining was carried out on transverse serial cryosections of muscles (10 µm). The muscles were sectioned at 10–12 different intervals along the length of the muscle, allowing the maximal CSA to be determined. For the assessment of tissue morphology and visualization of fibrosis and connective tissue, transverse sections of muscles were stained, respectively, with H&E and Sirius red for further examination under a light microscope. To assess central nucleation, three random areas were assessed in each section. The total number of fibres in these areas was counted and the number of centrally nucleated fibres was expressed as a percentage of the total number of fibres. For morphometric and fibre-type analyses, sections were air-dried, washed in phosphate-buffered saline (PBS) with 0.1% (v/v) Tween-20 (PBS-T) and stained for laminin (Dako, Z0097, Dako, Trappes, France) or for the different MyHC isoforms, with antibodies harvested from hybridoma cell lines obtained from the American Type Culture Collection (Manassas, VA, USA): BA-D5 (IgG2b, anti-MHCI), SC-71 (IgG1, anti-MHCIIa), BF-F3 (IgM, anti-MHCIIb) and 6H1 (IgM, anti-MHCIIX). The sections were incubated at room temperature for 1 h in a blocking solution [bovine serum albumin (BSA) 1%, sheep serum 1%, triton X-100 0.1%, sodium azid 0.001%]. Sections were then incubated at room temperature for 2 h with anti-MyHC-I (BA-D5, 2:3) and anti-MyHC-IIA (SC-71, 1:3). Sections were then incubated overnight at 4°C with anti-laminin (1:300) and anti-MyHC-IIb (BF-F3, 1:1) or anti-MyHC-IIX (6H1, 1:1). Sections were washed as before and secondary antibodies were applied for 1 h at a dilution of 1:400. Alexa 350 anti-mouse IgG2b, Cy3 anti-mouse IgG1, Alexa 647 anti-mouse IgM and Alexa 488 goat antirabbit were obtained from Vector Laboratories, Inc. (Burlingame, CA, USA). Metamorph software (Roper Scientific) was used to analyse the number, CSA and MyHC isoforms of fibres. For each muscle, the entire section was analysed.

For PABPN1 immunodetection, sections were blocked with 1% normal goat serum in 0.1 M PBS, 0,1% Triton X100 and incubated overnight at 4°C in primary antibody (a gift from Prof. Elmar Whale, Halle Germany) diluted to 1:500 in the same buffer. Slides were washed, incubated for 1 h with an anti-dystrophin antibody for fibre detection (NCL-Dys1 mouse monoclonal IgG2a, Novocastra), further incubated with respective secondary antibodies for 2 h at room temperature and stained with Hoechst to visualize nuclei. When necessary, sections were incubated in 1 M KCl, 30 mM HEPES, 65 mM PIPES, 10 mM EDTA, 2 mM $MgCl_2$, pH 6.9, for 1 h prior to the immunolabelling, to remove any soluble proteins.

Images were visualized using an Olympus BX60 microscope (Olympus Optical, Hamburg, Germany), digitalized using a CCD camera (Photometrics CoolSNAP fx; Roper Scientific, Tucson, AZ, USA) and analysed using MetaView image analysis system (Universal Imaging, Downington, PA, USA).

Proteasome peptidase activities

After dissection, TA from A17.1 and WT mice were homogenized for cytosolic extraction in a Polytron homogenizer (low setting, 3 s) using an ice-cold buffer containing: 50 mM Tris–HCl (pH 7.5), 250 mM sucrose, 5 mM $MgCl_2$, 2mM ATP, 1 mM DTT, 0.5 mM EDTA and 0.025% digitonin, as reported previously (Kisselev and Goldberg, 2005). The homogenate was centrifuged at 20000g for 15 min at 4°C. The pellet was discarded and the supernatant represents the cytosolic fraction (Kisselev and Goldberg, 2005). Protein quantification was made using the Bradford method (Pierce), with BSA as a standard. Peptidase activities of the proteasome were evaluated using appropriate fluorogenic substrates as described previously (Bulteau et al., 2001). Chymotrypsin-like (CT-like), trypsin-like (Tryp-like) and caspase-like (Casp-like) activities of the proteasome were assayed using the fluorogenic peptides LLVY-MCA (25 μM), RLR-MCA (40 μM) and LLE-NA (100 μM), respectively (Kisselev and Goldberg, 2005). The assay buffer was composed of 50 mM Tris–HCl (pH 7.5), 40 mM KCl, 5 mM $MgCl_2$, 1 mM DTT containing the appropriated peptide substrate. Enzymatic kinetics were carried out for 30 min at 37°C using 40 μg of cytosolic protein fractions in a temperature-controlled microplate fluorimetric reader (Fluostar Galaxy, bMG, Stuttgart, Germany). The excitation/emission wavelengths were 350/440 and 333/410 nm for aminomethylcoumarin and betanaphthylamine products. The rate of proteolysis was determined for each substrate as the mean slope by comparing the linear response of fluorescence with time. Reactions were performed in the presence (20 μM) and absence of the specific proteasome inhibitor N-Cbz-Leu-Leu-leucinal (MG132), to test the specificity of the activity measured.

Mitochondrial enzyme activity

All activities were determined at 30°C. Prior to analysis, cells were subjected to three cycles of freezing and thawing to lyse membranes. Enzyme activities were assessed using an Uvikon 940 spectrophotometer (Kontron Instruments Ltd, Watford, UK). Complex I activity was measured according to the method of Ragan et al. (Ragan et al., 1988). Complex II–III activity was measured according to the method of King et al. (King, 1967). Complex IV activity was measured according to the method of Wharton and Tzagoloff (Wharton and Tzagoloff, 1967). Citrate synthase (CS) activity was determined by the method of Shepherd and Garland (Shepherd and Garland, 1969). Enzyme activities are expressed as a ratio to CS (mitochondrial marker enzyme) to compensate for mitochondrial enrichment in the cell samples.

Western blotting

Muscle lysates were prepared by homogenizing tissue in RIPA solution (NaCl 0.15 M; HEPES 0.05 M; NP-40 1%; sodium dehoxycholate 0.5%; SDS 0.10%; EDTA 0.01 M) with protease inhibitor cocktail (Complete, Roche Diagnostics). Proteins were separated on 4–12% Bis–Tris gel (Invitrogen) and transferred onto a nitrocellulose membrane (Hybond ECL membrane; Amersham Biosciences), which was blocked by incubation in 5% milk in 0.1 M PBS, 0.1% Tween-20. Membrane was probed with primary antibodies raised against PABPN1 (gift from Pr. Elmar Wahle, Halle, Germany, 1:2000) or against GAPDH (Santa Cruz, 1:2000) as a loading control. The membrane was further incubated with HRP-conjugated antibodies (Jackson ImmunoResearch; 1:40000). Immunoreactive bands were detected with enhanced chemiluminescence reagent (ECL; Amersham Biosciences) and signals visualized by exposing the membrane to ECl Hyperfilm (Amersham Biosciences).

Statistical analysis

All data are presented as mean values ± standard error of the mean (SEM) (cohort size stated per experiment). All statistical analyses were performed using the Student t-test, the ANOVA one-way analysis of variance followed by the Newman–Keuls post-test, or X^2 analysis using GraphPad Prism (version 4.0b; GraphPad Software, San Diego CA, USA). A difference was considered to be significant at * P <0.05, ** P <0.01 or *** P <0.001.

Acknowledgments

We are grateful to Prof. Elmar Wahle (University of Halle, Germany) for providing the PABPN1 antibody.

We thank Claire Phillips and Nicola Sanderson for assistance with the breeding of the FvB and A17.1 mice, Linda Popplewell for the genotyping, Helen Foster and Alberto Malerba for sample collection, Anne Bigot and Christel Gentil for assistance with western blots. We also thank Julie Dumonceaux, Denis Furling, Arnaud Klein and the MSG for helpful discussions. Finally, we thank Bertrand Friguet for insightful comments and advice related to proteasome activity measurements.

Funding

This study was supported by grants from the European Commission ref. LSHM-CT-2005018675 – PolyALA, the Clinigene EC Network of Excellence (LSHB-CT-2006-018933), MYOAGE and MYORES Network of Excellence (contract FP7-LSHG-2007-B-223576 and FP6-511978), Wellcome Trust Senior Clinical fellowship, the Centre for Medical Systems Biology and the Muscular Dystrophy Association (68016), AFM (Association Française contre les Myopathies), INSERM, CNRS and Université Pierre et Marie Curie.

Reference List

Abu-Baker,A., Messaed,C., Laganiere,J., Gaspar,C., Brais,B., and Rouleau,G.A. (2003). Involvement of the ubiquitin-proteasome pathway and molecular chaperones in oculopharyngeal muscular dystrophy. Hum. Mol. Genet *12*, 2609-2623.

Apponi,L.H., Leung,S.W., Williams,K.R., Valentini,S.R., Corbett,A.H., and Pavlath,G.K. (2010). Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. Hum. Mol. Genet. *19*, 1058-1065.

Bao,Y.P., Cook,L.J., O'Donovan,D., Uyama,E., and Rubinsztein,D.C. (2002). Mammalian, yeast, bacterial, and chemical chaperones reduce aggregate formation and death in a cell model of oculopharyngeal muscular dystrophy. J. Biol. Chem. *277*, 12263-12269.

Blumen,S.C., Brais,B., Korczyn,A.D., Medinsky,S., Chapman,J., Asherov,A., Nisipeanu,P., Codere,F., Bouchard,J.P., Fardeau,M., Tome,F.M., and Rouleau,G.A. (1999). Homozygotes for oculopharyngeal muscular dystrophy have a severe form of the disease. Ann. Neurol. *46*, 115-118.

Bodine,S.C., Latres,E., Baumhueter,S., Lai,V.K., Nunez,L., Clarke,B.A., Poueymirou,W.T., Panaro,F.J., Na,E., Dharmarajan,K., Pan,Z.Q., Valenzuela,D.M., DeChiara,T.M., Stitt,T.N., Yancopoulos,G.D., and Glass,D.J. (2001). Identification of ubiquitin ligases required for skeletal muscle atrophy. Science *294*, 1704-1708.

Bottinelli,R. and Reggiani,C. (2000). Human skeletal muscle fibres: molecular and functional diversity. Prog. Biophys. Mol. Biol. *73*, 195-262.

Brais,B., Bouchard,J.P., Xie,Y.G., Rochefort,D.L., Chretien,N., Tome,F.M., Lafreniere,R.G., Rommens,J.M., Uyama,E., Nohira,O., Blumen,S., Korczyn,A.D., Heutink,P., Mathieu,J., Duranceau,A., Codere,F., Fardeau,M., and Rouleau,G.A. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet *18*, 164-167.

Brennan,K.J. and Hardeman,E.C. (1993). Quantitative analysis of the human alpha-skeletal actin gene in transgenic mice. J. Biol. Chem. *268*, 719-725.

Bruusgaard,J.C., Liestol,K., and Gundersen,K. (2006). Distribution of myonuclei and microtubules in live muscle fibers of young, middle-aged, and old mice. J. Appl. Physiol *100*, 2024-2030.

Bulteau,A.L., Lundberg,K.C., Humphries,K.M., Sadek,H.A., Szweda,P.A., Friguet,B., and Szweda,L.I. (2001). Oxidative modification and inactivation of the proteasome during coronary occlusion/reperfusion. J. Biol. Chem. *276*, 30057-30063.

Calado,A., Kutay,U., Kuhn,U., Wahle,E., and Carmo-Fonseca,M. (2000a). Deciphering the cellular pathway for transport of poly(A)-binding protein II. RNA. *6*, 245-256.

Calado,A., Tome,F.M., Brais,B., Rouleau,G.A., Kuhn,U., Wahle,E., and Carmo-Fonseca,M. (2000b). Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. Hum. Mol. Genet *9*, 2321-2328.

Calura,E., Cagnin,S., Raffaello,A., Laveder,P., Lanfranchi,G., and Romualdi,C. (2008). Meta-analysis of expression signatures of muscle atrophy: gene interaction networks in early and late stages. BMC. Genomics *9*, 630.

Catoire,H., Pasco,M.Y., Abu-Baker,A., Holbert,S., Tourette,C., Brais,B., Rouleau,G.A., Parker,J.A., and Neri,C. (2008). Sirtuin inhibition protects from the polyalanine muscular dystrophy protein PABPN1. Hum. Mol. Genet *17*, 2108-2117.

Chartier,A., Benoit,B., and Simonelig,M. (2006). A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. EMBO J *25*, 2253-2262.

Chartier,A., Raz,V., Sterrenburg,E., Verrips,C.T., van der Maarel,S.M., and Simonelig,M. (2009). Prevention of oculopharyngeal muscular dystrophy by muscular expression of Llama single-chain intrabodies in vivo. Hum. Mol. Genet. *18*, 1849-1859.

Chatterjee,S. and Price,B. (1991). Regression Analysis by Example. (New York: John Wiley and Sons).

Clarke,B.A., Drujan,D., Willis,M.S., Murphy,L.O., Corpina,R.A., Burova,E., Rakhilin,S.V., Stitt,T.N., Patterson,C., Latres,E., and Glass,D.J. (2007). The E3 Ligase MuRF1 degrades myosin heavy chain protein in dexamethasone-treated skeletal muscle. Cell Metab *6*, 376-385.

Cohen,S., Brault,J.J., Gygi,S.P., Glass,D.J., Valenzuela,D.M., Gartner,C., Latres,E., and Goldberg,A.L. (2009). During muscle atrophy, thick, but not thin, filament components are degraded by MuRF1-dependent ubiquitylation. J. Cell Biol. *185*, 1083-1095.

Corbeil-Girard,L.P., Klein,A.F., Sasseville,A.M., Lavoie,H., Dicaire,M.J., Saint-Denis,A., Page,M., Duranceau,A., Codere,F., Bouchard,J.P., Karpati,G., Rouleau,G.A., Massie,B., Langelier,Y., and Brais,B. (2005). PABPN1 overexpression leads to upregulation of genes encoding nuclear proteins that are sequestered in oculopharyngeal muscular dystrophy nuclear inclusions. Neurobiol. Dis. *18*, 551-567.

Davies,J.E., Sarkar,S., and Rubinsztein,D.C. (2006). Trehalose reduces aggregate formation and delays pathology in a transgenic mouse model of oculopharyngeal muscular dystrophy. Hum. Mol. Genet *15*, 23-31.

Davies,J.E., Sarkar,S., and Rubinsztein,D.C. (2008). Wild-type PABPN1 is anti-apoptotic and reduces toxicity of the oculopharyngeal muscular dystrophy mutation. Hum. Mol. Genet *17*, 1097-1108.

Davies,J.E., Wang,L., Garcia-Oroz,L., Cook,L.J., Vacher,C., O'Donovan,D.G., and Rubinsztein,D.C. (2005). Doxycycline attenuates and delays toxicity of the oculopharyngeal muscular dystrophy mutation in transgenic mice. Nat Med. *11*, 672-677.

Dennis,G., Jr., Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C., and Lempicki,R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, 3.

Dion,P., Shanmugam,V., Gaspar,C., Messaed,C., Meijer,I., Toulouse,A., Laganiere,J., Roussel,J., Rochefort,D., Laganiere,S., Allen,C., Karpati,G., Bouchard,J.P., Brais,B., and Rouleau,G.A. (2005). Transgenic expression of an expanded (GCG)13 repeat PABPN1 leads to weakness and coordination defects in mice. Neurobiol. Dis. *18*, 528-536.

Fan,X., Messaed,C., Dion,P., Laganiere,J., Brais,B., Karpati,G., and Rouleau,G.A. (2003). HnRNP A1 and A/B interaction with PABPN1 in oculopharyngeal muscular dystrophy. Can. J Neurol. Sci. *30*, 244-251.

Gilson,H., Schakman,O., Combaret,L., Lause,P., Grobet,L., Attaix,D., Ketelslegers,J.M., and Thissen,J.P. (2007). Myostatin gene deletion prevents glucocorticoid-induced muscle atrophy. Endocrinology *148*, 452-460.

Goeman,J.J., van de Geer,S.A., de,K.F., and van Houwelingen,H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. *20*, 93-99.

Gundersen,K. and Bruusgaard,J.C. (2008). Nuclear domains during muscle atrophy: nuclei lost or paradigm lost? J. Physiol *586*, 2675-2681.

Hino,H., Araki,K., Uyama,E., Takeya,M., Araki,M., Yoshinobu,K., Miike,K., Kawazoe,Y., Maeda,Y., Uchino,M., and Yamamura,K. (2004). Myopathy phenotype in transgenic mice expressing mutated PABPN1 as a model of oculopharyngeal muscular dystrophy. Hum. Mol. Genet *13*, 181-190.

Hirner,S., Krohne,C., Schuster,A., Hoffmann,S., Witt,S., Erber,R., Sticht,C., Gasch,A., Labeit,S., and Labeit,D. (2008). MuRF1-dependent regulation of systemic carbohydrate metabolism as revealed from transgenic mouse studies. J. Mol. Biol. *379*, 666-677.

Huang,d.W., Sherman,B.T., and Lempicki,R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. *4*, 44-57.

Jagoe,R.T., Lecker,S.H., Gomes,M., and Goldberg,A.L. (2002). Patterns of gene expression in atrophying skeletal muscles: response to food deprivation. FASEB J. *16*, 1697-1712.

Jelier,R., Schuemie,M.J., Veldhoven,A., Dorssers,L.C., Jenster,G., and Kors,J.A. (2008). Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. *9*, R96.

Kedar,V., McDonough,H., Arya,R., Li,H.H., Rockman,H.A., and Patterson,C. (2004). Muscle-specific RING finger 1 is a bona fide ubiquitin ligase that degrades cardiac troponin I. Proc. Natl. Acad. Sci. U. S. A *101*, 18135-18140.

Kim,Y.J., Noguchi,S., Hayashi,Y.K., Tsukahara,T., Shimizu,T., and Arahata,K. (2001). The product of an oculopharyngeal muscular dystrophy gene, poly(A)-binding protein 2, interacts with SKIP and stimulates muscle-specific gene expression. Hum. Mol. Genet *10*, 1129-1139.

King,T.E. (1967). Preparation of succinate cytochrome c reductase and cytochrome b-c1 particle and reconstruction of succinate cytochrome c reductase. Methods Enzymol. 446-451.

Kisselev,A.F. and Goldberg,A.L. (2005). Monitoring activity and inhibition of 26S proteasomes with fluorogenic peptide substrates. Methods Enzymol. *398*, 364-378.

Klein,A.F., Ebihara,M., Alexander,C., Dicaire,M.J., Sasseville,A.M., Langelier,Y., Rouleau,G.A., and Brais,B. (2008). PABPN1 polyalanine tract deletion and long expansions modify its aggregation pattern and expression. Exp. Cell Res. *314*, 1652-1666.

Kuhn,U., Gundel,M., Knoth,A., Kerwitz,Y., Rudel,S., and Wahle,E. (2009). Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J. Biol. Chem. *284*, 22803-22814.

Lecker,S.H., Jagoe,R.T., Gilbert,A., Gomes,M., Baracos,V., Bailey,J., Price,S.R., Mitch,W.E., and Goldberg,A.L. (2004). Multiple types of skeletal muscle atrophy involve a common program of changes in gene expression. FASEB J. *18*, 39-51.

Lemieux,C. and Bachand,F. (2009). Cotranscriptional recruitment of the nuclear poly(A)-binding protein Pab2 to nascent transcripts and association with translating mRNPs. Nucleic Acids Res. *37*, 3418-3430.

Li,P., Waters,R.E., Redfern,S.I., Zhang,M., Mao,L., Annex,B.H., and Yan,Z. (2007). Oxidative phenotype protects myofibers from pathological insults induced by chronic heart failure in mice. Am. J. Pathol. *170*, 599-608.

Marie-Josee,S.A., Caron,A.W., Bourget,L., Klein,A.F., Dicaire,M.J., Rouleau,G.A., Massie,B., Langelier,Y., and Brais,B. (2006). The dynamism of PABPN1 nuclear inclusions during the cell cycle. Neurobiol. Dis *23*, 621-629.

Messaed,C., Dion,P.A., Abu-Baker,A., Rochefort,D., Laganiere,J., Brais,B., and Rouleau,G.A. (2007). Soluble expanded PABPN1 promotes cell death in oculopharyngeal muscular dystrophy. Neurobiol. Dis *26*, 546-557.

Miniou,P., Tiziano,D., Frugier,T., Roblot,N., Le,M.M., and Melki,J. (1999). Gene targeting restricted to mouse striated muscle lineage. Nucleic Acids Res. *27*, e27.

Mouly,V., Aamiri,A., Bigot,A., Cooper,R.N., Di,D.S., Furling,D., Gidaro,T., Jacquemin,V., Mamchaoui,K., Negroni,E., Perie,S., Renault,V., Silva-Barbosa,S.D., and Butler-Browne,G.S. (2005). The mitotic clock in skeletal muscle regeneration, disease and cell mediated gene therapy. Acta Physiol Scand. *184*, 3-15.

Muqit,M.M., Larner,A.J., Sweeney,M.G., Sewry,C., Stinton,V.J., Davis,M.B., Healy,D.G., Payne,S.J., Chotai,K., Wood,N.W., and Lane,R.J. (2008). Multiple mitochondrial DNA deletions in monozygotic twins with OPMD. J. Neurol. Neurosurg. Psychiatry *79*, 68-71.

Nury,D., Doucet,C., and Coux,O. (2007). Roles and potential therapeutic targets of the ubiquitin proteasome system in muscle wasting. BMC. Biochem. *8 Suppl 1*, S7.

Orengo,J.P., Chambon,P., Metzger,D., Mosier,D.R., Snipes,G.J., and Cooper,T.A. (2008). Expanded CTG repeats within the DMPK 3' UTR causes severe skeletal muscle wasting in an inducible mouse model for myotonic dystrophy. Proc. Natl. Acad. Sci. U. S. A *105*, 2646-2651.

Pauzner,R., Blatt,I., Mouallem,M., Ben-David,E., Farfel,Z., and Sadeh,M. (1991). Mitochondrial abnormalities in oculopharyngeal muscular dystrophy. Muscle Nerve *14*, 947-952.

Pearson,K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine *2*, 559-572.

Perie,S., Mamchaoui,K., Mouly,V., Blot,S., Bouazza,B., Thornell,L.E., St Guily,J.L., and Butler-Browne,G. (2006). Premature proliferative arrest of cricopharyngeal myoblasts in oculo-pharyngeal muscular dystrophy: Therapeutic perspectives of autologous myoblast transplantation. Neuromuscul Disord *16*, 770-781.

Ragan,C.I., Wilson,M.Y., and Darley-Usman,V.M. (1988). Subfractionation of mitochondria and isolation of proteins of oxidative phosphorylation. In Mitochondria: a practical approach, V.M.Darley, D.Rickwood, and M.T.Wilson, eds. (Oxford: IRL Press), pp. 79-113.

Sacheck,J.M., Hyatt,J.P., Raffaello,A., Jagoe,R.T., Roy,R.R., Edgerton,V.R., Lecker,S.H., and Goldberg,A.L. (2007). Rapid disuse and denervation atrophy involve transcriptional changes similar to those of muscle wasting during systemic diseases. FASEB J. *21*, 140-155.

Sandri,M. (2008). Signaling in muscle atrophy and hypertrophy. Physiology. (Bethesda.) *23*, 160-170.

Schroder,J.M., Krabbe,B., and Weis,J. (1995). Oculopharyngeal muscular dystrophy: clinical and morphological follow-up study reveals mitochondrial alterations and unique nuclear inclusions in a severe autosomal recessive type. Neuropathol. Appl. Neurobiol. *21*, 68-73.

Shepherd,J.A. and Garland,P.B. (1969). Citrate synthase activity from rat liver. Methods Enzymol. *13*, 11-19.

Smyth,G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Smyth,G.K. and Speed,T. (2003). Normalization of cDNA microarray data. Methods *31*, 265-273.

Tavanez,J.P., Bengoechea,R., Berciano,M.T., Lafarga,M., Carmo-Fonseca,M., and Enguita,F.J. (2009). Hsp70 chaperones and type I PRMTs are sequestered at intranuclear inclusions caused by polyalanine expansions in PABPN1. PLoS. One. *4*, e6418.

Tavanez,J.P., Calado,P., Braga,J., Lafarga,M., and Carmo-Fonseca,M. (2005). In vivo aggregation properties of the nuclear poly(A)-binding protein PABPN1. RNA. *11*, 752-762.

Tome,F.M., Chateau,D., Helbling-Leclerc,A., and Fardeau,M. (1997). Morphological changes in muscle fibers in oculopharyngeal muscular dystrophy. Neuromuscul Disord *7 Suppl 1*, S63-S69.

Toriumi,K., Oma,Y., Kino,Y., Futai,E., Sasagawa,N., and Ishiura,S. (2008). Expression of polyalanine stretches induces mitochondrial dysfunction. J. Neurosci. Res. *86*, 1529-1537.

Toriumi,K., Oma,Y., Mimoto,A., Futai,E., Sasagawa,N., Turk,B., and Ishiura,S. (2009). Polyalanine tracts directly induce the release of cytochrome c, independently of the mitochondrial permeability transition pore, leading to apoptosis. Genes Cells *14*, 751-757.

Uyama,E., Hino,H., Araki,K., Takeya,M., Uchino,M., and Yamamura,K. (2005). Animal model of oculopharyngeal muscular dystrophy. Acta Myol. *24*, 84-88.

Vazeille,E., Codran,A., Claustre,A., Averous,J., Listrat,A., Bechet,D., Taillandier,D., Dardevet,D., Attaix,D., and Combaret,L. (2008). The ubiquitin-proteasome and the mitochondria-associated apoptotic pathways are sequentially downregulated during recovery after immobilization-induced muscle atrophy. Am. J. Physiol Endocrinol. Metab *295*, E1181-E1190.

Vignaud,A., Fougerousse,F., Mouisel,E., Guerchet,N., Hourde,C., Bacou,F., Butler-Browne,G.S., Chatonnet,A., and Ferry,A. (2008). Genetic inactivation of acetylcholinesterase causes functional and structural impairment of mouse soleus muscles. Cell Tissue Res. *333*, 289-296.

Vignaud,A., Ramond,F., Hourde,C., Keller,A., Butler-Browne,G., and Ferry,A. (2007). Diabetes provides an unfavorable environment for muscle mass and function after muscle injury in mice. Pathobiology *74*, 291-300.

Wahle,E. (1991). A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. Cell *66*, 759-768.

Wahle,E. (1995). Poly(A) tail length control is caused by termination of processive synthesis. J. Biol. Chem. *270*, 2800-2808.

Wharton,D.C. and Tzagoloff,A. (1967). Cytochrome oxidase from beef heart mitochondria. Methods Enzymol. *10*, 245-250.

Wirtschafter,J.D., Ferrington,D.A., and McLoon,L.K. (2004). Continuous remodeling of adult extraocular muscles as an explanation for selective craniofacial vulnerability in oculopharyngeal muscular dystrophy. J. Neuroophthalmol. *24*, 62-67.

Yu,Z., Li,P., Zhang,M., Hannink,M., Stamler,J.S., and Yan,Z. (2008). Fiber type-specific nitric oxide protects oxidative myofibers against cachectic stimuli. PLoS. One. *3*, e2086.

Zhang,Y., Marcillat,O., Giulivi,C., Ernster,L., and Davies,K.J. (1990). The oxidative inactivation of mitochondrial electron transport chain components and ATPase. J. Biol. Chem. *265*, 16330-16336.

APPENDIX

Generation of expPABPN1 construct and viral vectors

The expanded PABPN1 cDNA was obtained from Dr. Michael Antoniou (Department of Medical and Molecular Genetics, King's College London). The cDNA sequence was cloned into a pDD-derived AAV plasmid under the control of the CAGGs promoter. To produce the rAAV2/8-CAG-expPABPN1, HEK293T cells were transfected with the expression plasmid and the helper plasmids pAdDF6 and pAAV5E18-VD2/8 (James Wilson, University of Pennsylvania, Philadelphia, PA) using calcium phosphate precipitation. Cell pellets were harvested and lysed in 50 mmol/l TrisHCl, 150 mmol/l NaCl. Lysates were clarified by centrifugation at 6,700 rpm for 20 minutes and passed through a 0.45-μm filter. Cell lysates were layered on an iodixanol gradient (Sigma-Aldrich, Poole, UK) and centrifuged at 60,000 rpm for 90 minutes. The 40% iodixanol layer containing the viral particles was isolated, concentrated with phosphate-buffered saline (PBS), 5 mmol/l MgCl2, 12.5 mmol/l KCl (PBSMK), through an Amicon Ultra-15 100 kd (Millipore, UK). The number of vector genomes was determined relative to a plasmid DNA standard using Dot blot hybridisation.

Administration of rAAV

Eight-week-old FvB mice were anaesthetised by intraperitoneal injection of 3.75 ml/g body weight of premixed (1:1) Hypnorm/Hypnovel (Hypnorm: Janssen Pharmaceutical, Belgium; Hypnovel: Hoffmann-La Roche Ltd, Switzerland). The lower hindlimbs were shaved and the TA muscles injected with 1x1012 vector genomes or rAAV2/8-CAG-expPABPN1 diluted in injectable saline (Sigma-Aldrich). Muscle contractile properties and histological assessments of injected tibialis anterior (TA) muscles were performed three months following administration of rAAV.

Western blotting

Muscle lysates were prepared by homogenising tissue in RIPA solution (NaCl 0.15M; Hepes 0.05M; NP-40 1%; Sodium dehoxycholate 0.5%; SDS 0.10%; EDTA 0.01M) with protease inhibitor cocktail (Complete, Roche Diagnostics) and phosphatase inhibitor cocktail (20mM NaF, 10mM b-glycérophosphate, 5mM Na-pyrophsphate, and 1mM Naorthovanadate). Proteins were separated on 4-12% Bis-Tris gel (Invitrogen) and transferred onto a nitrocellulose membrane (Hybond ECL membrane; Amersham Biosciences), which was blocked by incubation in 5% BSA in 0.1M TBS, 0.1% Tween-20. Membrane was probed with primary antibodies raised against PABPN1 (gift from Pr. Elmar Wahle, Halle, Germany, 1:2000), MuRF1 (Abcam; Ab-4125; 1:500), Foxo3A (Abcam; Ab-12162; 1:1000), Akt (Cell Signaling Technology; 9272; 1:1000), Phospho-Akt-Ser473 (Cell Signaling Technology; 9271; 1:1000). The membrane was further incubated with HRP-conjugated antibodies (Jackson ImmunoResearch; 1:40000). Immunoreactive bands were detected with enhanced chemiluminescence reagent (ECL; Amersham Biosciences) and signals visualised by exposing the membrane to ECl Hyperfilm (Amersham Biosciences).

Immunohistochemistry

For Foxo3A immunodetection, sections were air dried, fixed with 4% paraformaldehyde, incubated 10 minutes with NH4Cl 50mM, blocked with 4% BSA in 0.1 M PBS, 0,1% Triton X100 and incubated overnight at 4°C in primary antibody (Foxo3A; Cell Signaling Technology; 2497) diluted 1:100 in the same buffer. Slides were washed, incubated for one hour with anti-dystrophin antibody for fibre detection (NCL-Dys1 mouse monoclonal IgG2a, Novocastra), further incubated with respective secondary antibodies for 2 hours at room temperature and stained with Hoechst to visualise nuclei.

Table S1 - Complete list of the most significant A17.1 deregulated biological processes GO terms, including sub-categories.

ID	GO term	P-value	Genes	Deregulated genes	
GO:0051169	**Nuclear transport**	**1.35E-08**	**94**	**46**	**(49%)**
GO:0006913	nucleocytoplasmic transport	2.27E-07	93	46	(49%)
GO:0046822	regulation of nucleocytoplasmic transport	2.27E-07	19	11	(58%)
GO:0051170	nuclear import	2.27E-07	60	28	(47%)
GO:0046823	negative regulation of nucleocytoplasmic transport	2.27E-07	7	6	(86%)
GO:0051168	nuclear export	2.27E-07	22	11	(50%)
GO:0046824	positive regulation of nucleocytoplasmic transport	8.37E-07	7	2	(29%)
GO:0009056	**Catabolic process**	**1.41E-08**	**872**	**361**	**(41%)**
GO:0009894	regulation of catabolic process	1.12E-03	33	13	(39%)
GO:0044248	cellular catabolic process	1.87E-03	749	323	(43%)
GO:0009057	macromolecule catabolic process	2.60E-03	583	271	(46%)
GO:0016052	carbohydrate catabolic process	5.41E-03	61	21	(34%)
GO:0016042	lipid catabolic process	2.60E-02	116	28	(24%)
GO:0015031	**Protein transport**	**1.60E-08**	**591**	**250**	**(42%)**
GO:0006886	intracellular protein transport	2.27E-07	274	124	(45%)
GO:0017038	protein import	2.27E-07	68	32	(47%)
GO:0051224	negative regulation of protein transport	2.28E-07	10	6	(60%)
GO:0051223	regulation of protein transport	2.32E-07	34	10	(29%)
GO:0042953	lipoprotein transport	4.28E-07	7	1	(14%)
GO:0051222	positive regulation of protein transport	9.19E-07	13	2	(15%)
GO:0009306	protein secretion	2.59E-04	32	4	(13%)
GO:0045859	**Regulation of protein kinase activity**	**1.68E-08**	**132**	**50**	**(38%)**
GO:0006469	negative regulation of protein kinase activity	2.93E-04	36	15	(42%)
GO:0000079	regulation of cyclin-dependent protein kinase activity	3.44E-04	7	5	(71%)
GO:0043405	regulation of MAP kinase activity	7.75E-04	65	21	(32%)
GO:0045860	positive regulation of protein kinase activity	2.52E-03	91	32	(35%)
GO:0006796	**Phosphate metabolic process**	**1.69E-08**	**804**	**313**	**(39%)**
GO:0045937	positive regulation of phosphate metabolic process	2.80E-04	44	15	(34%)
GO:0045936	negative regulation of phosphate metabolic process	8.86E-04	21	9	(43%)
GO:0019220	regulation of phosphate metabolic process	1.03E-03	93	33	(35%)
GO:0016311	dephosphorylation	9.67E-03	116	57	(49%)
GO:0006072	glycerol-3-phosphate metabolic process	1.03E-02	7	4	(57%)
GO:0006950	**Response to stress**	**1.77E-08**	**955**	**289**	**(30%)**
GO:0001666	response to hypoxia	6.27E-03	26	9	(35%)
GO:0006970	response to osmotic stress	1.30E-02	15	8	(53%)
GO:0009408	response to heat	1.47E-02	26	10	(38%)
GO:0006979	response to oxidative stress	1.69E-02	53	23	(43%)
GO:0006986	response to unfolded protein	2.11E-02	29	8	(28%)
GO:0006974	response to DNA damage stimulus	2.81E-02	240	104	(43%)
GO:0033554	cellular response to stress	2.82E-02	304	128	(42%)
GO:0006952	defense response	3.02E-02	379	69	(18%)
GO:0009611	response to wounding	3.75E-02	315	71	(23%)
GO:0006457	**Protein folding**	**1.87E-08**	**107**	**47**	**(44%)**
GO:0006458	'de novo' protein folding	1.51E-05	13	7	(54%)
GO:0006397	**mRNA processing**	**1.90E-08**	**219**	**111**	**(51%)**
GO:0000398	nuclear mRNA splicing, via spliceosome	2.27E-07	27	11	(41%)
GO:0050684	regulation of mRNA processing	3.71E-07	4	3	(75%)

ID	GO term	P-value	Genes	Deregulated genes	
GO:0031124	mRNA 3'-end processing	2.96E-06	8	2	(25%)
GO:0007049	**Cell cycle**	**1.93E-08**	**615**	**197**	**(32%)**
GO:0000278	mitotic cell cycle	1.37E-02	232	69	(30%)
GO:0022402	cell cycle process	1.69E-02	338	104	(31%)
GO:0045786	negative regulation of cell cycle	4.87E-02	83	34	(41%)
GO:0050790	**Regulation of catalytic activity**	**1.96E-08**	**331**	**114**	**(34%)**
GO:0043086	negative regulation of catalytic activity	1.12E-03	65	22	(34%)
GO:0051338	regulation of transferase activity	2.17E-03	140	52	(37%)
GO:0043085	positive regulation of catalytic activity	1.39E-02	186	59	(32%)
GO:0006915	**Apoptosis**	**2.03E-08**	**647**	**225**	**(35%)**
GO:0043066	negative regulation of apoptosis	1.87E-02	189	69	(37%)
GO:0042981	regulation of apoptosis	2.49E-02	426	145	(34%)
GO:0043065	positive regulation of apoptosis	2.58E-02	183	64	(35%)
GO:0051402	neuron apoptosis	3.30E-02	67	22	(33%)
GO:0051276	**Chromosome organization and biogenesis**	**2.33E-08**	**319**	**135**	**(42%)**
GO:0006325	establishment or maintenance of chromatin architecture	2.27E-07	250	119	(48%)
GO:0000819	sister chromatid segregation	3.78E-07	17	3	(18%)
GO:0030261	chromosome condensation	1.19E-06	15	2	(13%)
GO:0070192	chromosome organization involved in meiosis	2.98E-06	15	4	(27%)
GO:0032200	telomere organization	1.41E-05	19	6	(32%)
GO:0033044	regulation of chromosome organization	5.43E-04	9	2	(22%)
GO:0007517	**Muscle development**	**2.49E-08**	**179**	**73**	**(41%)**
GO:0048747	muscle fiber development	2.27E-07	55	21	(38%)
GO:0048644	muscle morphogenesis	2.32E-07	7	2	(29%)
GO:0048634	regulation of muscle development	9.28E-07	26	13	(50%)
GO:0048635	negative regulation of muscle development	6.74E-06	7	4	(57%)
GO:0007525	somatic muscle development	2.32E-05	3	2	(67%)
GO:0048636	positive regulation of muscle development	1.19E-04	2	1	(50%)
GO:0009628	**Response to abiotic stimulus**	**2.60E-08**	**185**	**60**	**(32%)**
GO:0006970	response to osmotic stress	1.30E-02	15	8	(53%)
GO:0009314	response to radiation	2.41E-02	106	36	(34%)
GO:0009266	response to temperature stimulus	2.89E-02	43	14	(33%)
GO:0009612	response to mechanical stimulus	3.88E-02	24	2	(8%)
GO:0007005	**Mitochondrion organization**	**2.68E-08**	**60**	**24**	**(40%)**
GO:0008637	Apoptotic mitochondrial changes	4.55E-08	21	6	(29%)
GO:0006461	**Protein complex assembly**	**2.89E-08**	**164**	**66**	**(40%)**
GO:0043623	cellular protein complex assembly	2.27E-07	115	48	(42%)
GO:0031334	positive regulation of protein complex assembly	2.68E-07	11	7	(64%)
GO:0051259	protein oligomerization	5.26E-07	38	14	(37%)
GO:0043254	regulation of protein complex assembly	5.41E-07	42	20	(48%)
GO:0031333	negative regulation of protein complex assembly	4.72E-06	21	9	(43%)
GO:0010608	**posttranscriptional regulation of gene expression**	**2.27E-07**	**95**	**45**	**(47%)**
GO:0006417	regulation of translation	2.27E-07	113	41	(36%)
GO:0031647	regulation of protein stability	3.69E-07	46	14	(30%)
GO:0016441	posttranscriptional gene silencing	4.35E-07	6	3	(50%)
GO:0043487	regulation of RNA stability	9.63E-03	18	3	(17%)
GO:0006511	**ubiquitin-dependent protein catabolic process**	**2.27E-07**	**451**	**215**	**(48%)**
GO:0043161	proteasomal ubiquitin-dependent protein catabolic process	2.27E-07	20	11	(55%)
GO:0042787	protein ubiquitination during protein catabolic process	2.32E-07	4	2	(50%)

ID	GO term	P-value	Genes	Deregulated genes	
GO:0016567	**Protein ubiquitination**	**1.88E-05**	**53**	**26**	**(49%)**
GO:0031398	positive regulation of protein ubiquitination	2.27E-07	3	2	(67%)
GO:0000209	protein polyubiquitination	2.27E-07	13	9	(69%)
GO:0031396	regulation of protein ubiquitination	2.27E-07	8	4	(50%)
GO:0042787	protein ubiquitination during protein catabolic process	2.32E-07	4	2	(50%)
GO:0051865	protein autoubiquitination	3.24E-05	1	1	(100%)
GO:0006412	**translation**	**9.69E-03**	**273**	**139**	**(51%)**
GO:0017148	negative regulation of translation	2.27E-07	21	8	(38%)
GO:0006417	regulation of translation	2.27E-07	113	41	(36%)
GO:0006413	translational initiation	2.27E-07	52	19	(37%)
GO:0006414	translational elongation	2.27E-07	23	10	(43%)
GO:0006418	tRNA aminoacylation for protein translation	2.27E-07	69	22	(32%)
GO:0006415	translational termination	1.37E-06	8	3	(38%)
GO:0045727	positive regulation of translation	4.08E-04	7	2	(29%)
GO:0042692	**muscle cell differentiation**	**9.80E-03**	**75**	**31**	**(41%)**
GO:0014902	myotube differentiation	2.27E-07	22	6	(27%)
GO:0051146	striated muscle cell differentiation	2.59E-07	72	28	(39%)
GO:0055001	muscle cell development	3.16E-07	38	18	(47%)
GO:0045445	myoblast differentiation	7.58E-06	28	4	(14%)
GO:0051147	regulation of muscle cell differentiation	1.07E-05	35	7	(20%)
GO:0051149	positive regulation of muscle cell differentiation	1.10E-05	15	4	(27%)
GO:0051145	smooth muscle cell differentiation	7.99E-05	22	5	(23%)
GO:0051148	negative regulation of muscle cell differentiation	1.61E-04	13	2	(15%)
GO:0048666	**neuron development**	**1.31E-02**	**276**	**83**	**(30%)**

Table S2 - A) List of the 1,679 overlapping deregulated genes recognized by Anni 2.0; **B)** List of the 481 deregulated genes highly associated with the terms 'muscle atrophy' or 'skeletal muscle atrophy'; **C)** List of the 163 genes showing a progression profile; **D)** List of the 63 selected progressive genes related to muscle atrophy.

(excel file can be retrived from *http://hmg.oxfordjournals.org/content/suppl/2010/03/03/ddq098.DC1/ddq098_supp_table_2.xls***)**

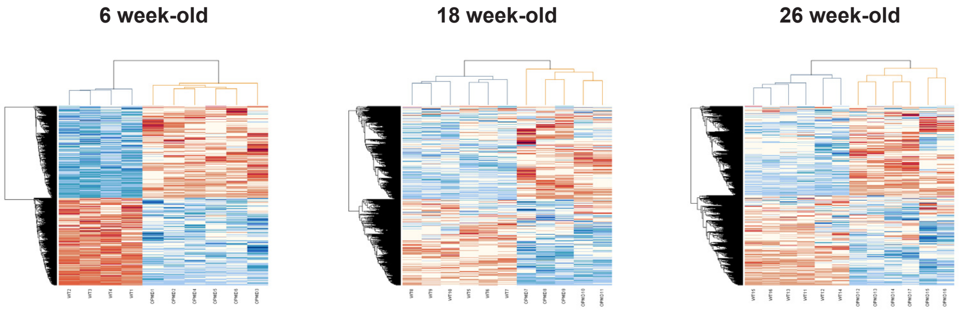
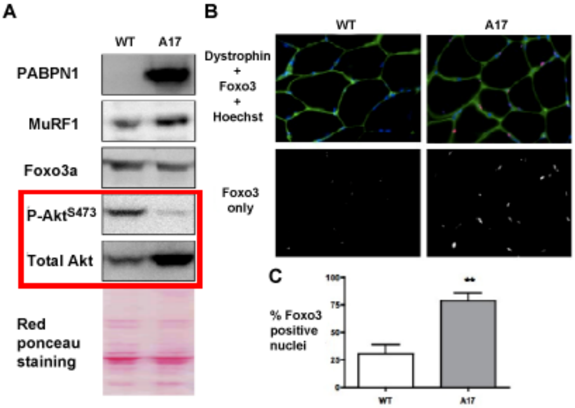


Figure S1 - Clustergrams (heat maps) for each time point of the transcriptomic analysis.

Figure S2 – A) Muscle extracts from 26-week-old A17.1 and WT mice were immunoblotted with the indicated antibody. **B)** Immunofluorescence staining for total Foxo3A (red), dystrophin (green) and Hoechst staining (blue) on WT and A17.1 TA muscle sections, magnification x400. **C)** The percentage of nuclei containing a positive total Foxo3A staining was determined on skeletal muscle (TA) cryosections from 26-week-old A17.1 mice (∗∗ p<0.01).



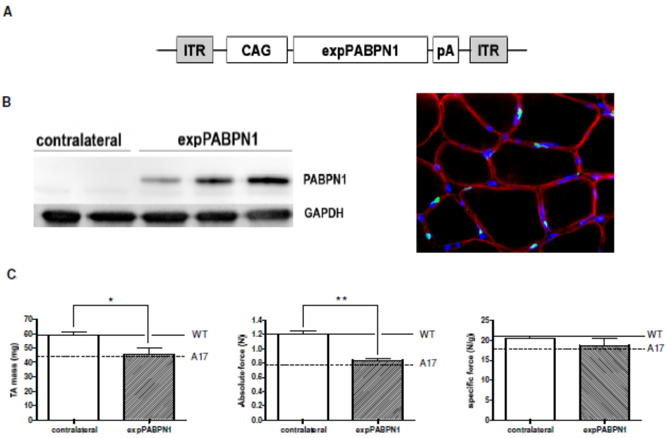


Figure S3 - Measurements of the weight and functional performance of skeletal muscle in WT mice injected with rAAV-CAG-expPABPN1 compared to the uninjected contralateral leg (n=3 per group). A) Diagram of the rAAV2/8-CAG-expPABPN1 construct. **B)** Western-blot to confirm the overexpression of expPABPN1 and immunostaining to confirm the localization in nuclei of injected muscle fibers (expPABPN1 in green, dystrophin staining in red, nuclei in blue). **C)** Measurements of the muscle mass and the maximal absolute force of both TA for each mouse: injected leg (expPABPN1) and contralateral uninjected leg (contralateral) (n=3; ∗∗ P<0.01; ∗ P<0.05). As a comparison we have indicated the corresponding muscle mass and maximal absolute force of the WT (plain line) and A17.1 mice (dash line).

CHAPTER TWO

Deregulation of the ubiquitin-proteasome system is the predominant molecular pathology in OPMD animal models and patients

Seyed Yahya Anvar[1], Peter A.C. ‘t Hoen[1], Andrea Venema[1], Barbara van der Sluijs[2], Baziel van Engelen[2], Marc Snoeck[3], John Vissing[4], Capucine Trollet[5,6], George Dickson[5], Aymeric Chartier[7], Martine Simonelig[7], Gert-Jan B van Ommen[1], Silvère M. van der Maarel[5] and Vered Raz[1,*]

Oculopharyngeal muscular dystrophy (OPMD) is a late-onset progressive muscle disorder caused by a poly-alanine expansion mutation in PABPN1. The molecular mechanisms that regulate disease onset and progression are largely unknown. In order to identify molecular pathways that are consistently associated with OPMD, we performed an integrated high-throughput transcriptome study in affected muscles of OPMD animal models and patients. The ubiquitin-proteasome system (UPS) was found as the most consistently and significantly deregulated pathway across species. We could correlate the association of the UPS deregulated genes with stages of disease progression. The expression trend of a subset of these genes is age-associated and therefore marks the late onset of the disease, and a second group with expression trends relating to disease-progression. We demonstrate a correlation between expression trends and entrapment in PABPN1 insoluble aggregates of deregulated E3 ligases. We also show that manipulations of proteasome and immunoproteasome activity specifically affect the accumulation and aggregation of mutant PABPN1. We suggest that the natural decrease in proteasome expression and its activity during muscle aging contributes to the onset of the disease.

---

1 Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. **2** Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands. **3** Department of Anaesthesia, Canisius-Wilhelmina Hospital, Nijmegen, the Netherlands. **4** Neuromuscular Research Unit and Department of Neurology, University of Copenhagen, Rigs Hospitalet, Denmark. **5** School of Biological Sciences, Royal Holloway, University of London, Egham, UK. **6** INSERM U974, UMR 7215 CNRS, Institut de Myologie, UM 76 Université Pierre et Marie Curie, Paris, France. **7** Institut de Genetique Humaine, CNRS UPR1142, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France.

* To whom correspondence should be addressed at: v.raz@lumc.nl

Skeletal Muscle, 2011,
Apr 4;1(1):15; doi:10.1186/2044-5040-1-15

BACKGROUND

Oculopharyngeal muscular dystrophy (OPMD) is a late-onset progressive muscle disorder for which the underlying molecular mechanisms are largely unknown. This autosomal dominant muscular dystrophy has an estimated prevalence of 1 in 100,000 worldwide (Fan and Rouleau, 2003). A higher prevalence has been reported in the Jewish Caucasian and French-Canadian populations (1 in 600 and 1 in 1000, respectively) (Blumen et al., 2009; Brais et al., 1995). OPMD is caused by expansion of a homopolymeric alanine (Ala) stretch at the N-terminus of the Poly(A) Binding Protein Nuclear 1 (PABPN1) by 2-7 additional Ala residues (Brais et al., 1998). Although PABPN1 is ubiquitously expressed, the clinical and pathological features of OPMD are restricted to a subset of skeletal muscles, causing progressive *ptosis*, *dysphagia*, and limb muscle weakness. In affected muscles, the expanded PABPN1 (expPABPN1) accumulates in intranuclear inclusions (INI) (Tome and Fardeau, 1980). Animal models for OPMD were generated in *Drosophila,* mouse and *C. elegans* with a muscle-specific expression of expPABPN1 (Chartier et al., 2006; Davies et al., 2005; Catoire et al., 2008). These models recapitulate INI formation and progressive muscle weakness in OPMD, and a correlation between INI formation and muscle weakness has been reported (Chartier et al., 2006; Davies et al., 2005; Catoire et al., 2008). In these OPMD models protein disaggregation approaches attenuate muscle symptoms (Davies et al., 2006; Catoire et al., 2008; Chartier et al., 2009). So far, however, the molecular mechanisms that are associated with OPMD onset and progression are not known. Previously, we preformed transcriptome analysis on skeletal muscles from a mouse model of OPMD and found massive gene deregulation, which was reflected by a broad spectrum of altered cellular pathways (Trollet et al., 2010). We found an association of transcriptional changes with muscle atrophy (Trollet et al., 2010). Muscle atrophy was recently reported in homozygous OPMD patients (Blumen et al., 2009). However, the vast majority of OPMD patients are heterozygous and muscle atrophy is not common pathological characteristic of the disease in its early stages. Importantly, a mouse model with low and constitutive expPABPN1 expression exhibits minor muscle defects without muscle atrophy (Hino et al., 2004). Hino et al. (2004) suggested that the extent of muscle symptoms caused by expPABPN1 depends on the expression level. Therefore, it is not known whether the massive transcriptional changes in affected muscles of the A17.1 OPMD model (Trollet et al., 2010) are due to the high over-expression of expPABPN1 or that they are common with transcriptional changes in OPMD patients.

We have generated microarrays of OPMD carriers at pre-symptomatic and symptomatic stages. Since OPMD is categorized as a rare disorder in Western countries, limited patient material is an obstacle in reaching conclusive results. Therefore, we performed a cross-species transcriptome study by integrating transcriptome data from *Drosophila* and mouse models and heterozygous OPMD patients. We hypothesized that OPMD-associated molecular mechanisms would be consistently deregulated across species. As bioinformatics analyses of gene expression are biased by the computational approaches (Ioannidis et al., 2009), here we integrated three computational methods to obtain a higher degree of confidence and reproducibility. The ubiquitin-proteasome system (UPS) was identified as the most significant and consistent OPMD-deregulated pathway across species.

RESULTS

Genome-wide expression profiles from the *Drosophila* and mouse OPMD models (Chartier et al., 2006; Trollet et al., 2010) were integrated with the expression profiles of heterozygous OPMD carriers (datasets are described in **Table S1** and **Table S2**). Genes that are differentially expressed between OPMD and controls (OPMD-deregulated) were identified using limma model in R

Table 1 - Deregulation of ubiquitin-proteasome system (UPS) in OPMD in *Drosophila,* human and mouse. For *Drosophila* and mouse *P*-values are derived from the combined analysis of the three time points using global test, where age was included as a confounder in the model.

	Drosophila	Mouse	Human
Literature Analysis			
Ubiquitination	# 2	# 1	# 1
GO Categories			
Ubiquitin-dependent Protein Catabolic Process	2.81E-04	2.27E-07	1.22E-03
Protein Ubiquitination	7.57E-03	1.88E-05	9.24E-04
Proteasomal Protein Catabolic Process	6.51E-03	2.23E-07	1.86E-03
KEGG Pathways			
Ubiquitin Mediated Proteolysis	2.03E-03	8.25E-08	1.52E-03
Proteasome	2.15E-04	1.37E-07	9.27E-03

(Smyth, 2004). To identify the most prominent and consistent feature across all species, comparative pathway analysis was performed using three computational methods (**Figure S1**). In literature-aided analyses (Jelier et al., 2008), the term 'ubiquitination' was found to be the most strongly associated biomedical concept with OPMD-deregulated genes (**Table 1** and **Table S3**). A regression-based analysis using global test (GT) (Goeman et al., 2004), and an enrichment method using DAVID (Dennis, Jr. et al., 2003; Huang et al., 2009) revealed highly significant deregulation of ubiquitin-proteasome system (UPS)-related GO (Gene Ontology) categories and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways across species (**Table 1**).

To evaluate the level of concordance between the animal models and OPMD patients, gene overlap between the OPMD-deregulated UPS genes was determined. Homologous genes were annotated using the HomoloGene and Inparanoid databases (see Methods). In total, 16%, 32% and 25% of the genes annotated to the UPS were identified as OPMD-deregulated in *Drosophila*, mouse and human, respectively (**Figure 1A**). More than half of the OPMD-deregulated genes in *Drosophila* (59%) overlapped with their mouse or human homologous genes, and close to half (45% and 51%) overlapped between mouse and human genes, respectively (**Figure 1A**). The similarity of deregulation direction across species was demonstrated for 14 genes, for which probes were found in all organisms (**Figure 1B**). Similar transcriptional changes were found for 13 homologous genes in mouse and human datasets. Among those, 8 genes showed similar changes in *Drosophila*. These results show the consistent UPS deregulation in OPMD.

To validate the microarray analyses, quantitative RT-PCR (Q-PCR) was performed on 19 OPMD-deregulated UPS genes from mouse. Genes were selected based on *P*-value and >1.3 fold change criteria. For 17/19 genes (89%), Q-PCR results confirmed the results of the microarray analyses (**Figure S2**). This demonstrates the reproducibility and validity of the microarray statistical analyses.

In the A17.1 mouse model, muscle atrophy is more prominent in fast glycolytic fibers (quadriceps) as compared with slow oxidative fibers (soleus) (Trollet et al., 2010). Since muscle atrophy is regulated by the UPS (Cao et al., 2005; Bodine et al., 2001; Sandri, 2008), we analyzed the muscle-

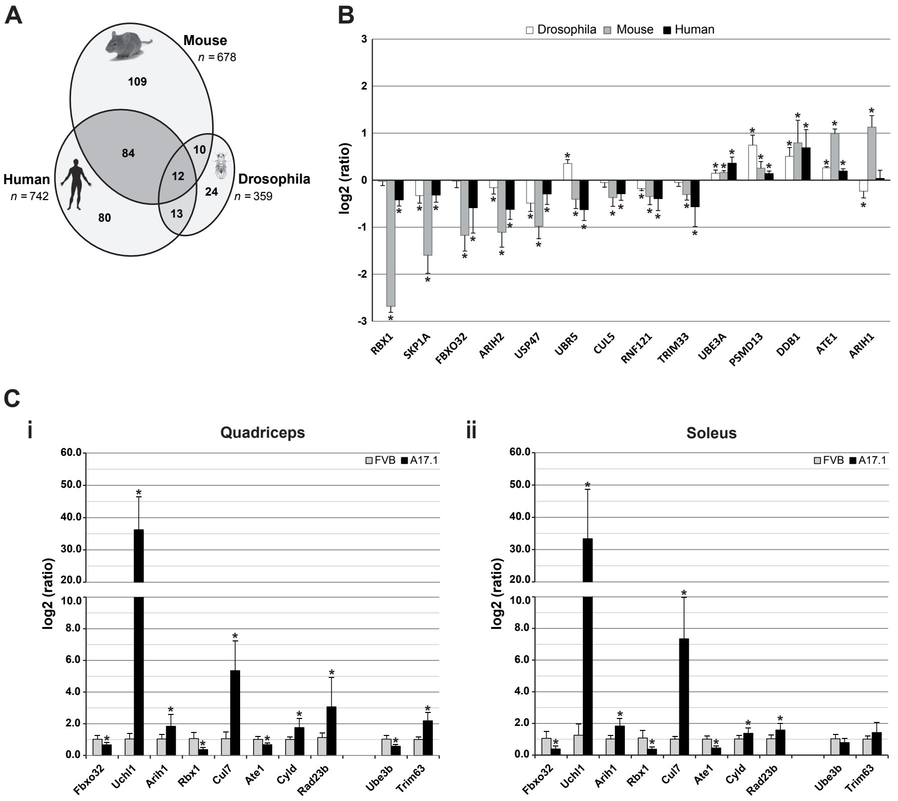


Figure 1 - Cross species deregulation of ubiquitin-proteasome in OPMD. A) Venn-diagram displaying the overlap in OPMD-deregulated genes in UPS across species. In mouse and *Drosophila*, OPMD-deregulated genes should be consistently deregulated in at least two time points. The total number of genes in UPS is indicated in italics. The list of OPMD-deregulated UPS genes is in Additional File 1. **B)** Transcriptional changes of selected genes in UPS in different organisms. Histograms display the log2(ratio) of the measured expression values in *Drosophila* (white bars), mouse (gray bars), and human (black bars). Significant changes with the adjusted $P < 0.05$ are indicated by $*$. **C)** RT Q-PCR validation of selected deregulated genes in UPS was carried out on quadriceps (**i**) and soleus (**ii**) muscles of 6 week-old mice. Histograms show the measured expression values for A17.1 and FVB mice using Q-PCR. Significant changes of measured expression values of A17.1 mice as compared to FVB with the $P < 0.05$ are indicated by $*$.

---

type specific expression of 10 OPMD-deregulated UPS genes in order to identify a correlation with muscle atrophy. Q-PCR was performed on RNA isolated from quadriceps and soleus of 6 week-old A17.1 and control (FVB) mice. The majority (8 out of 10) of genes showed no fiber-type specificity (**Figure 1C**). Only the deregulation of *Trim63* (Trollet et al., 2010) and *Ube3b* were specific to fast glycolytic fibers (**Figure 1C**). This suggests that the majority of OPMD-deregulated UPS genes are not associated with muscle atrophy in the A17.1 mouse.

The UPS involves an enzymatic cascade of ubiquitination and degradation steps. The ubiquitination steps start with ubiquitin activation, which requires the ubiquitin-activating enzyme (E1) and ubiquitin (Ub). This process results in the binding of Ub to the E2-conjugating enzyme. In a subsequent step the target protein is ubiquitinated with Ub-E2 and E3-ligase complexes, which ensures target specificity. Poly-ubiquitinated proteins are subjected to degradation. This

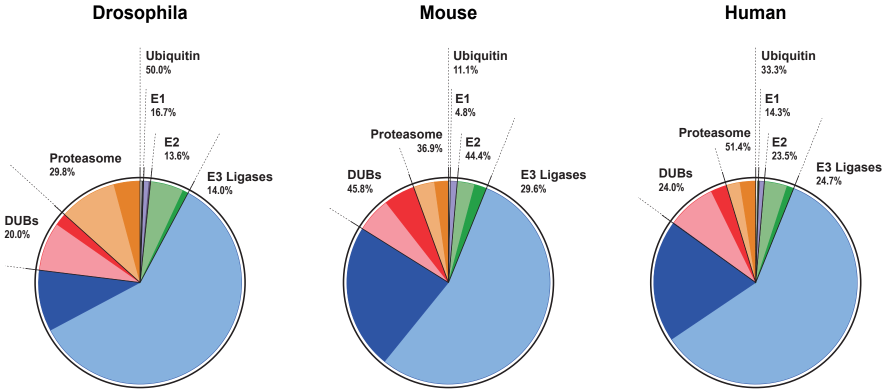


Figure 2 - Pie charts show the relative distribution of the UPS units (light colors) and OPMD-deregulated genes (dark colors) per organism. Numbers indicate the percentage of OPMD-deregulation.

step is employed by the deubiquitinating enzymes (DUBs) and the proteasome (Reyes-Turcu et al., 2009; Finley, 2009). Deregulation of genes involved in the ubiquitin activation step was not found to be consistent between OPMD and the models (**Figure 2** and **Table 2**). Ubiquitin up-regulation was previously reported in a non-muscle cell model for OPMD (Abu-Baker et al., 2003). Our study identified only one ubiquitin-encoding gene to be up regulated in mouse and human genomes, but these deregulated genes were not consistent across species. The E2-conjugating enzymes were significantly deregulated in *Drosophila* and mouse genomes, whereas in humans, the *P*-value for these enzymes was not significant. This suggests a weak association of E2 deregulation with OPMD (**Figure 2** and **Table 2**). In contrast, consistent deregulation was found for E3-ligases, DUBs, and proteasome (**Figure 2** and **Table 2**). The significance of this strong association was further evaluated by gene-overlap of homologous genes in human and mouse (**Table 2**). The gene overlap between mouse and human was found to be significant for all these three UPS components (*P*-values are 6.64E-08 for E3-ligases, 1.37E-02 for DUBs and 1.70E-02 for the proteasome). Overall, this analysis demonstrates consistent deregulation of E3-ligases, DUBs and proteasome across species.

OPMD is characterized by a late onset and a slow progression of muscle weaknesses (VICTOR et al., 1962; Brais et al., 1998). Progressive muscle weakness has also been reported in the mouse model (Davies et al., 2005). In 6 week-old mice symptoms were not detected, while muscle weakness was present in 18 week-old mice and was more pronounced by 26 weeks (Davies et al., 2005). If changes in expression levels are associated with disease onset and progression, a correlation between age and expression levels should be expected. A linear regression model was applied to the mouse UPS genes at three time points in order to identify genes that their expression trends are progressively changed. 80% of the OPMD-deregulated UPS genes show a progressive trend, which is age-associated (N=171/217, **Figure S3A,** examples for progressive expression trends are shown in **Figure 3Ai**). To identify genes with expression trends that are specific to the disease a regression model that combines age and disease features was applied. In 30% of the age-associated OPMD-deregulated UPS genes (N=50, **Figure S3B)** the progression trends significantly (*P*-value<0.05) differed between A17.1 and the wild-type (WT) controls (examples for progressive expression trends are shown in **Figure 3Bi**). The genes with disease-specific progression can be used to mark disease progression and could contribute to disease onset and progression. The

Table 2 - The distribution of OPMD-deregulated genes in UPS functional units and protein degradation categories. The number of annotated genes per unit, the percentage of OPMD-deregulated (D.E.) genes and *P*-values are indicated per organism. For *Drosophila* and mouse statistics is generated in combined datasets from three time points. The overlap in OPMD-deregulated genes between human and mouse and the percentage of deregulated genes in human D.E. genes are indicated. Protein degradation machineries are depicted by ¥.

	Drosophila		**Mouse**		**Human**		**Overlap mouse vs. human**	
	% D.E. Genes	P-Value (FDR)	% D.E. Genes	P-Value (FDR)	% D.E. Genes	P-Value (FDR)	# D.E. Genes	% D.E. Genes
Ubiquitin	50.00	4.18E-05	11.11	1.28E-01	33.33	1.19E-01	0	00.00
E1 Ubiquitin Activation	16.67	1.24E-01	04.76	7.29E-02	14.29	7.94E-02	0	00.00
E2 Ubiquitin Conjugation	13.64	9.19E-06	44.42	1.51E-08	23.53	7.31E-02	3	37.50
E3 Ubiquitin Ligase	13.99	1.92E-04	29.58	1.64E-08	24.74	4.35E-03	69	47.59
Deubiquitination (DUB)	20.00	1.63E-05	45.83	1.48E-08	24.00	3.15E-02	13	72.22
¥ Proteasome	29.79	2.15E-04	36.94	1.37E-07	51.35	9.27E-03	11	57.90
¥ Autophagy	25.00	1.07E-03	30.77	8.13E-08	18.75	1.37E-02	1	16.67
¥ Lysosome	5.00	1.64E-02	25.33	6.06E-03	24.68	1.54E-02	6	33.33

group of genes whose expression changes with age independent from the disease, however, may contribute to the late onset of the disease.

The vast majority of OPMD-deregulated UPS genes, which exhibit progressive expression profiles encode for E3-ligases (**Figure S3**). Expression trends for selected E3-ligases are presented in Figure 3. Confirmation of the analysis in mouse was carried out on the human homologues (**Figure 3**). The age-associated expression trends were similar between A17.1 and WT in mouse and between controls and expPABPN1 carriers at pre-symptomatic and symptomatic stages in human (**Figure 3A**). The progression trends did not significantly differ between genotypes (*P*-value > 0.05). In contrast, for those genes with expression trends associated with age and disease the expression trends of controls significantly differed from those of OPMD subjects (**Figure 3B**, *P*-value <0.05). Validation of progression analysis was performed by Q-PCR analysis of RNA from 6 and 26 week-old mice (**Figure 3C**). The Q-PCR results demonstrate the reproducibility and validity of the microarray progression analysis.

In the progression analysis some differences between human and mouse were noted. The progression of *Trim63* is mouse-specific, whereas the expression of the human *TRIM63* is not age-associated or OPMD-deregulated (**Figure 3B)**. *Asb11* is down regulated in mouse while it is up regulated in human (**Figure 3A**). The expression trend of *Socs4* in mouse is negative while in human it is positive (**Figure 3B**). These discrepancies could reflect differences between the two organisms or between the heterozygous and the high over-expression situation.

Expression of expPABPN1 leads to INI formation in affected muscles (Davies et al., 2005; Trollet et al., 2010). Previous studies have demonstrated that ubiquitin and proteasome proteins co-localize with INI in affected muscles (Calado et al., 2000) and in non-muscle cells (Abu-Baker et al., 2003; Tavanez et al., 2005). Since INI formation is a hallmark of OPMD, we studied whether the expression profiles of OPMD-deregulated E3-ligases correlate with their entrapment with expPABPN1 in INI. Co-localization was analyzed with an immunofluorescence procedure in C2C12 myotubes expressing expPABPN1 fused to yellow fluorescent protein (YFP). From the

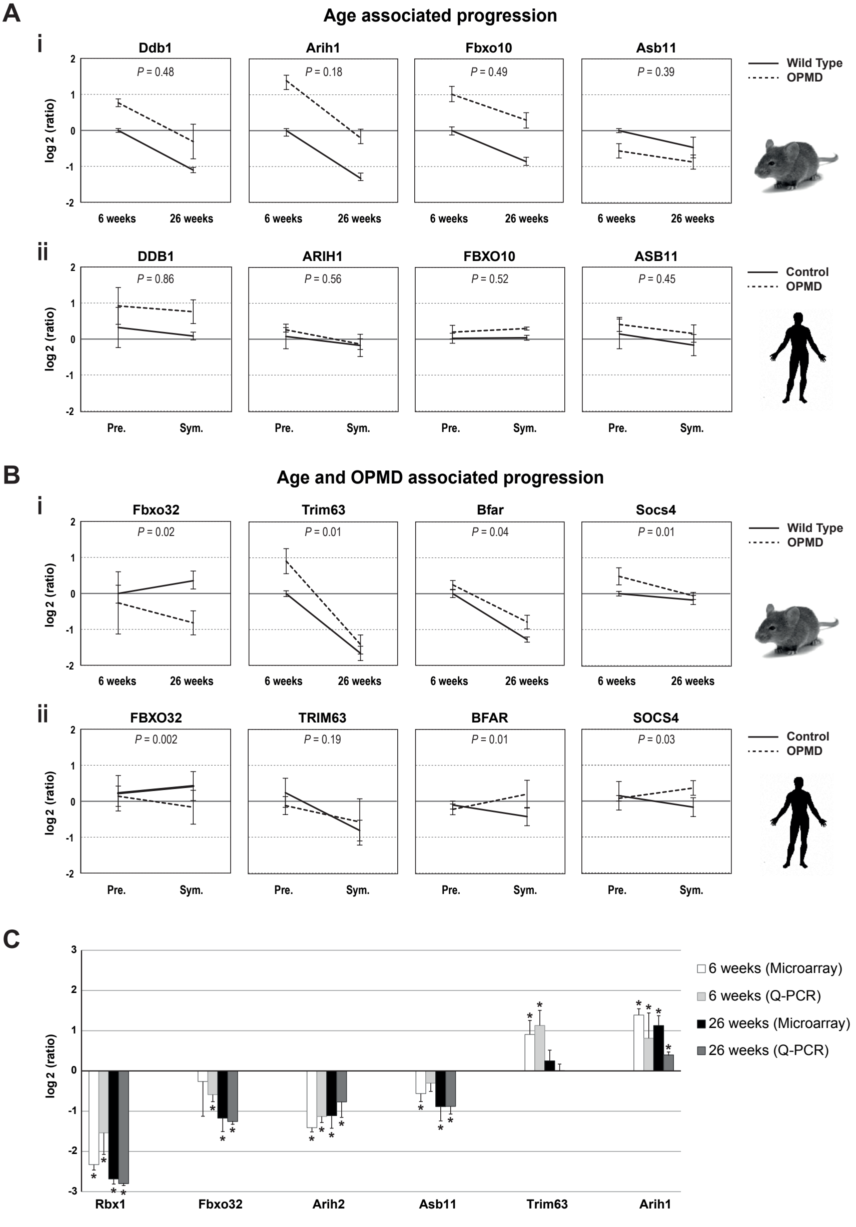


Figure 3 - Progressive changes in UPS gene expression. Progression trends for selected genes in mouse (**i**) and human (**ii**). Expression values were normalized to 6 weeks-old WT in mouse, and to young healthy controls (19 years-old in average) in human. *P*-values demonstrate the significance of differences in expression trends between controls and OPMD samples. **A)** The age-associated progression trend is indicated by *P*-value >0.05. **B)** The genotype-specific progression trend is indicated by *P*-values <0.05. SD represents variations in mouse (6 weeks N=5 and 26 weeks N=6) and in human (expPABPN1 carriers N=4 and controls N=5). **C)** RT Q-PCR validation of selected deregulated genes in UPS was carried out on skeletal muscles of 6 week-old and 26 week-old mice. Histograms show the log2(ratio) of the measured expression values using microarray and Q-PCR. Significant changes with the $P < 0.05$ are indicated by $*$.

E3-ligases encoding genes that showed an association with disease onset or progression (**Figure 3**), five were selected for co-localization studies using specific antibodies recognizing single proteins at the appropriate molecular weights. All 5 proteins showed nuclear localization in myotubes and co-localized with expPABPN1 in INI (**Figures 4**). Arih1, Asb11 and Ddb1 co-localized with all sizes of INI structures (**Figure 4A**) while the co-localization of Trim63 and Fbxo32 proteins were only evident for larger INI structures (**Figure 4B**, highlighted in boxes). This suggests a correlation between changes in expression trends and temporal entrapment in INI.

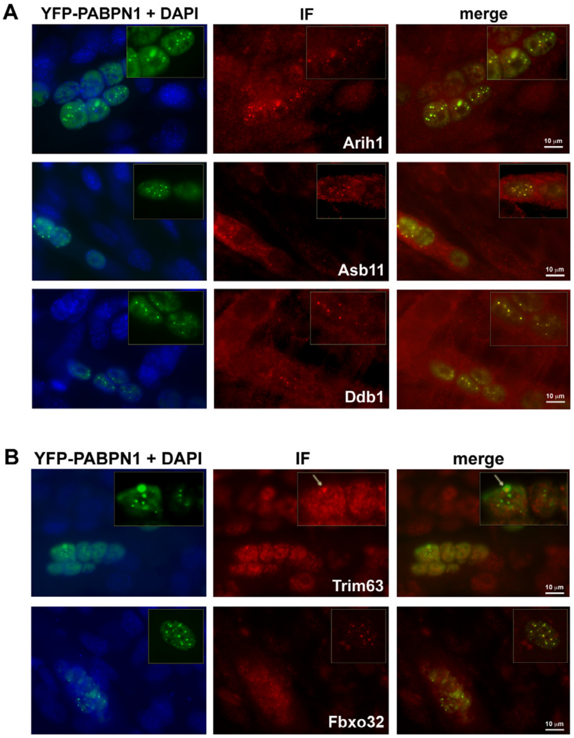


Figure 4 - Co-localization of selected E3 ligases with INI in C2C12 myotubes expressing YFP-Ala16PABPN1. Immunostaining of E3-ligases was visualized with Alexa-594 secondary antibodies. Co-localization with expPABPN1 in myotubes is demonstrated in the merge image. A 2.5X magnification of nuclei containing expPABPN1 aggregates is highlighted in a box. **A)** Arih1, Asb11 and Ddb1 E3 ligases show consistent co-localization with aggregated YFP-Ala16-PABPN1. **B)** Trim63 and Fbxo32 E3 ligases show progressively more co-localization with YFP-Ala16-PABPN1 as INI size increases. Scale bar is 10µm.

The proteasome is composed of core and regulatory subunits. Genes encoding for the proteasome core subunit were prominently down regulated in mouse and human (66% and 75%, respectively), while no preference in deregulation direction was found for the regulatory subunit (**Figure 5A** and **Table S5**). Down-regulation of the proteasome could affect protein degradation and, hence, protein accumulation. In C2C12 myoblasts that were treated with low concentrations (5 µM) of the proteasome inhibitor MG132, the accumulation of expPABPN1 was significantly higher as compared with mock-treated cells (**Figure 5B**). Similarly, treatment with the DUB inhibitor, PR619, also caused expPABPN1 accumulation (**Figure 5B**). High nuclear accumulation of expPABPN1, which accompanies INI formation, was consistently measured in MG132 treated cells using a cell-based intensity fluorescence quantification assay (**Figure 5C**). Thus, reduced proteasome and DUB activities in muscle cells promoted expPABPN1 accumulation and INI formation in muscle cells. However, expPABPN1 accumulation stimulated by proteasome inhibition is not specific to muscle cells (Abu-Baker et al., 2003).

In addition to the proteasome, the lysosome and the autophagy machineries can also facilitate protein catabolism. To evaluate whether one of these machineries could also regulate expPABPN1 protein accumulation the significance of deregulation in OPMD was analyzed. Overall, deregulation of lysosome and autophagy were not consistent across species. The lysosome KEGG pathway was evaluated as significantly deregulated in OPMD across species by GT but not by DAVID analysis (**Table 2**). However, in the literature-aided analysis, only a low level of association was found between OPMD-deregulated genes and lysosome in *Drosophila* and human

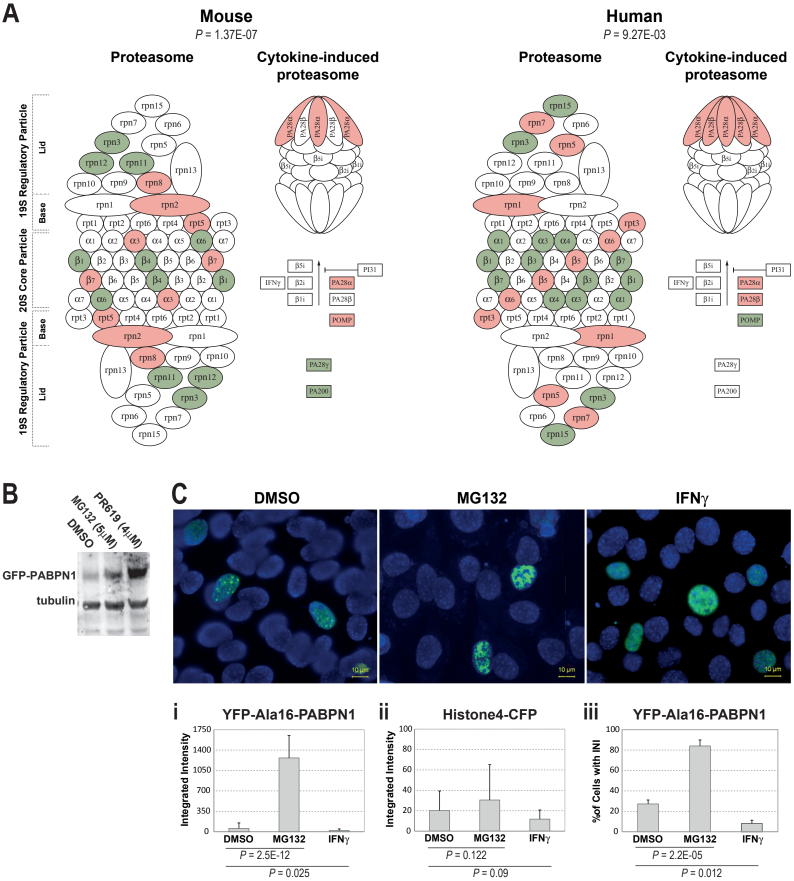


Figure 5 - The effect of altered proteasome activity on expPABPN1 accumulation and aggregation. **A)** Substantial deregulation of proteasome and immunoproteasome encoding genes in mouse and human. Down-regulation (green) is more pronounced in the core subunit of the proteasome. Immunoproteasome shows consistent up-regulation (red) in both organisms. **B)** Western blot analysis of YFP-Ala16-PABPN1 transfected C2C12 cells that were treated with 5μM MG132 or 5nM PR619. Control cells were treated with DMSO. **C)** Images show YFP-Ala16-PABPN1 localization in C2C12 after mock-treatment (DMSO), 5 μM MG132 or 5U/ml IFNγ. Scale bar equals 10 μm. Histograms show the integrated intensity of YFP-Ala16-PABPN1 (**i**) or Histone4-CFP (control) (**ii**), and the percentage of cells with INI in YFP-Ala16-PABPN1 expressing cells (**iii**). Averages represent 509, 773 and 476 cells for DMSO, MG132 and IFNγ, respectively. Significant difference between treatments is reflected by *P*-values.

(ranked at positions 196 and 789, respectively), while no association was found in mice. Similarly, the autophagy KEGG pathway was significant across species based on GT but not on DAVID analysis (**Table 2**). In the literature-aided analysis, autophagy was ranked 12 in mice, but a lower priority (ranked 136 and 154) was found in *Drosophila* and humans, respectively. Furthermore, the OPMD-deregulated gene overlap between mouse and human were not significant for either lysosome or autophagy pathways (*P*-values: 5.37E-01 and 3.70E-01, respectively). This is in sharp contrast to the consistent proteasome deregulation found across species. This indicates that, from the protein degradation pathways, only proteasome deregulation is consistently associated with OPMD across species. From this analysis we cannot exclude lysosome or autophagy deregulation in OPMD, but the lack of consistency across species and in three bioinformatics analyses suggests a smaller contribution as compared with the proteasome.

In contrast to the down regulation of genes in the core-subunit, the expression of genes encoding for the immunoproteasome subunit (the cytokine-induced proteasome) was consistently elevated in OPMD (**Figure 5A**). The immunoproteasome was initially identified in cells of the immune system after cytokine induction, which is involved in MHC-class-I antigen presentation (Kloetzel and Ossendorp, 2004). However, the accumulation of cytokine-induced proteasome proteins was also found in aging skeletal muscle cells (Ferrington et al., 2005). Treatment of C2C12 myoblasts with IFNγ, an inducer of immunoproteasome activity (Osna et al., 2003), led to a significant reduction in nuclear expPABPN1 accumulation (**Figure 5C** and **5Ci**) and INI formation (**Figure 5Ciii**). In contrast to expPABPN1, accumulation of Histone4, which is also a nuclear protein, was not significantly affected by manipulation of proteasome activity (**Figure 5Cii**). This suggests that the accumulation of expPABPN1, but not of Histone4, is receptive to the level of proteasome and immunoproteasome activity. Together, our results demonstrate that the UPS degradation machinery regulates expPABPN1 accumulation.

DISCUSSION

UPS is a cellular regulator of homeostasis and is involved in a wide spectrum of human diseases including cancer, neurodegenerative disorders and diabetes (Hoeller and Dikic, 2009; Liu et al., 2000; Combaret et al., 2009; Taillandier et al., 2004; Ciechanover and Brundin, 2003). Deregulation of UPS has been reported for myotonic dystrophy type 1 (Vignaud et al., 2010) and muscle atrophy in mice (Cao et al., 2005; Bodine et al., 2001; Sandri, 2008). In addition, altered UPS activity has been associated with muscle ageing (Combaret et al., 2009; Lee et al., 1999). Together these studies suggest that muscle cell function is tightly regulated by the UPS. In this study, we identified the UPS as the most consistently and significantly deregulated cellular machinery in OPMD animal models and patients. Transcriptome studies in non-muscle cells expressing expPABPN1 did not reveal substantial and predominant deregulation of UPS genes (Corbeil-Girard et al., 2005). This indicates that the effect of expPABPN1 on UPS deregulation is specific to muscle cells. Since PABPN1 is ubiquitously expressed in every cell but the phenotype is limited to muscle cells this suggests that UPS deregulation confers the muscle-specific pathogenesis of OPMD.

From six UPS components, only E3-ligases, DUBs and the proteasome were found to be consistently and prominently deregulated in OPMD across species. Relevant to OPMD proteasome activity is reduced during muscle aging (Combaret et al., 2009; Lee et al., 1999; Ferrington et al., 2005), and is associated with transcriptional deregulation of proteasomal genes (Lee et al., 1999). In the analysis of expression trends the expression of 89% of the OPMD-deregulated proteasome genes were found to be age-associated. This suggests that the natural decrease in proteasome expression during muscle aging can contribute to the late onset of the disease. Our analysis revealed that the core subunit of the proteasome is the only UPS subunit that was consistently down regulated which can cause reduced activity of the proteasome machinery. In a recent study, we found that expression of expPABPN1 in myotubes leads to down-regulation of proteasome-encoding genes, and causing the accumulation of expPABPN1 protein (unpublished data). However, proteasome regulation of expPABPN1 accumulation and INI formation is not specific to muscle cells (Abu-Baker et al., 2003). Since in patients INI are formed only in muscle cells this suggests that proteasome down-regulation during muscle aging triggers expPABPN1 accumulation. In turn, accumulation of expPABPN1 leads to extensive proteasome down-regulation in OPMD (**Figure 6**). This feed forward model could justify the muscle-specific INI formation and the late onset in OPMD.

Hypothesizing that changes in expression levels could reflect pathological changes in disease sta-

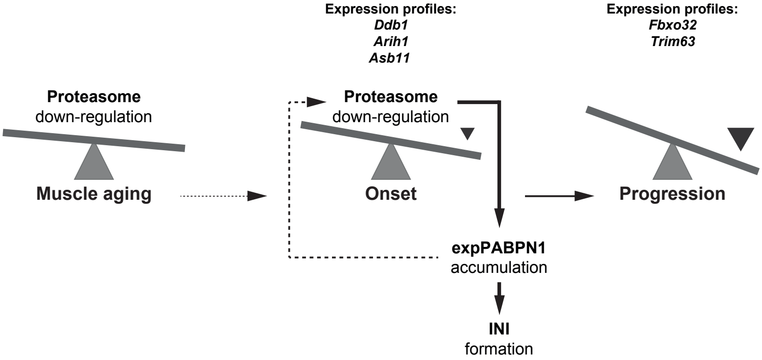


Figure 6 - A model for the involvement of UPS in OPMD disease pathology. In muscle, age-associated proteasome down regulation affects expPABPN1 protein accumulation. Elevated expPABPN1 accumulation affects proteasome deregulation during disease onset. Expression profiles of E3-ligases can be sued to separate disease onset from progression.

tus we have studied the correlation between transcriptional changes of OPMD-deregulated UPS genes and age. Noticeably, the expression of the vast majority of the OPMD-deregulated UPS genes is progressed during normal muscle aging. This suggests that transcriptional changes of these genes are associated with disease onset. The expression trends of a subset of these genes showed disease-specific progression. Among those, *Trim63* and *Fbxo32* exhibited disease progression in mouse. Both genes are known for regulating muscle atrophy in mice (Cao et al., 2005; Bodine et al., 2001; Sandri, 2008). In the OPMD mouse model, muscle atrophy in exhibited only in fast muscle glycolytic fibres and *Trim63* expression correlates with muscle atrophy in A17.1 (Trollet et al., 2010). However, the majority of the OPMD-deregulated UPS genes did not show fibre-type specific expression. This could suggest that UPS deregulation in OPMD has a broader pathological effect than muscle atrophy. Indeed, in affected muscles of OPMD patients, atrophy may be evident only at a later stage of disease progression. Although a high degree of consistency between expression trends in mice and human was found for the majority of the genes analyzed in this study. *Trim63* deregulation and progression is probably mouse-specific, as OPMD-deregulation or progression was not found in human. *Fbxo32,* however, was consistently deregulated in both organisms and, therefore, can be a candidate for regulating disease progression and muscle atrophy in human. After mining the NCBI dataset for tissue-specific expression (Unigene Hs.352183, Build No. 228 released 2010), *Asb11* was noted for its specific expression in skeletal muscles**.** Since *Asb11* is consistently OPMD-deregulated in human and mouse, and its expression trend is associated with disease onset it could represent a relevant candidate for functional genomic studies. This shows that cross-species transcriptome and progression analyses can be used to identify target molecules for future studies.

OPMD is characterized by INI formation. The role of INIs in disease pathogenesis is unknown. Previous studies have shown that many genes whose expression is deregulated by expPABPN1 are found to be co-localized in INI (Corbeil-Girard et al., 2005). Components of the proteasome, which is OPMD-deregulated, also co-localize in INI (Abu-Baker et al., 2003; Tavanez et al., 2005). We also found that many of the OPMD deregulated E3-ligases are entrapped in INI. Moreover, we demonstrate a correlation between temporal changes in expression levels and sequential entrapment in INI. Together these studies suggest that entrapment in INI could lead to transcriptional deregulation. It is possible that protein entrapment in INI affects gene expression through a compensatory mechanism resulting in altered transcriptional profiles.

CONCLUSIONS

In this study, we combined expression datasets from three organisms and disease models with different bioinformatics analyses in a single study. This allowed us to identify with high confidence the UPS as the most predominantly deregulated cellular pathway in OPMD. This approach differs from most microarray studies where results are derived from a single computational analysis performed on a single organism. We show that with this combined bioinformatics approach the list of deregulated pathways can be prioritized with high confidence. This approach can facilitate studies with complex biological situations and massive gene deregulation such as late onset disorders and rare-diseases.

The most significant and novel finding in this study is the substantial and cross-species consistent deregulation of the ubiquitin-proteasome system (UPS) in OPMD. We propose that protein entrapment in PABPN1 aggregates is associated with a substantial transcriptional deregulation of the UPS that, in turn leads to disruption of homeostasis in skeletal muscles. By taking advantage of the detailed analysis of gene expression trends and muscle- expression, we predict that candidate genes can be selected for functional genomic studies which ultimately lead to the identification of OPMD pathogenesis.

METHODS

Generation of microarray datasets

Drosophila and mouse microarray datasets have previously been published (Chartier et al., 2009; Trollet et al., 2010). Human quadriceps muscle samples were collected with the needle or by an open surgical procedure from OPMD patients and family members as well as from anonymous age-matching healthy individuals that gave informed consent. The presence of expansion mutation in PABPN1 in OPMD patients and pre-symptomatic was determined with sequencing. Bergstrom needle biopsies from the (pre)symptomatic patients were approved by the ethical committee. Total RNA was extracted from skeletal muscles using RNA Bee (Amsbio) according to the manufacturer's instructions. RNA integration number (RIN) was determined with RNA 6000 Nano (Agilent Technologies). RNA with RIN >7 were used for subsequent steps. RNA labeling was performed with the Illumina® TotalPrep RNA Amplification kit (Ambion) according to the manufacturer's protocol, and subsequently was hybridized to Illumina Human v3 Bead arrays. The generated microarray datasets are deposited and publicly available at GEO repository. GEO accession numbers for mouse and human microarray datasets are GSE26604 and GSE26605, respectively.

Data processing and statistical analysis

Microarray measurements were normalized using the quantile method (Smyth and Speed, 2003). Each organism and time point was normalized separately. The quality of the data was assessed by principal component analysis.

For *Drosophila* and mouse, genes differentially expressed between OPMD and control subjects were identified at each time point by applying a hierarchical linear model using the limma package in R (Smyth, 2004). Human subjects were grouped into healthy, pre-symptomatic and symptomatic subjects. *P*-value cut-offs of 0.05 after multiple-testing correction using the method of Benjamini and Hochberg (False Discovery Rate (FDR) were applied to the *Drosophila* and mouse samples and, due to higher inter-individual variation, a nominal *P*-value cut-off of 0.05 was used for human samples) were used. This resulted in lists of OPMD-deregulated genes for each time point and organism. Probe annotation was done using the indac (*Drosophila*), illuminaMousev-

1BeadID (mouse), and illuminaHumanv3BeadID (human) R packages. The OPMD significantly deregulated genes in the UPS from human and mouse datasets are listed in Additional File 1.

Pathway analyses

Global test (GT) (Goeman et al., 2004) was used to identify significant associations between GO categories or KEGG pathways and OPMD, while including age as a confounder (*Drosophila* and mouse only). Gene sets with multiple testing adjusted (Holm's method) *P-value* < 0.05 were selected as significant. DAVID, a functional annotation clustering tool (Dennis, Jr. et al., 2003; Huang et al., 2009), was applied on a list of OPMD-deregulated genes and pathway redundancy was removed by clustering similar GO categories and pathways. In addition, biomedical concepts that are associated with OPMD-deregulated genes were identified using literature-aided mapping tool, Anni 2.0 (Jelier et al., 2008). The procedure was performed for each organism separately. Cross-species analyses were carried out on a group of homologous genes. *Drosophila* homologues of mouse and human genes were annotated using HomoloGene (*http://www.ncbi.nlm.nih.gov/* homologene) and Inparanoid *(http://inparanoid.sbc.su.se)* online databases. Integration of three time-points in *Drosophila* and mouse (**Table S1**) were used to identify OPMD-deregulated pathways across species (**Figure 1**). A recent annotation of E3 ligases (Li et al., 2008) was used to identify OPMD-deregulated E3 ligases. The annotation for all other UPS components is extracted from KEGG. Since the annotation for genes encoding for lysosome is not available in R packages, we have extracted the annotation from KEGG website and integrated it into our pathway analyses.

Progression studies

For testing the significance of the association of expression trends of OPMD-deregulated genes with age, using limma model in R (Smyth, 2004), a linear regression model ($expression \sim \alpha OPMD + \beta AGE + \delta(OPMD \text{ x } AGE) + \varepsilon$) was applied on combined datasets from 6 and 26 weeks old mice. Age-associated changes were identified as those with β significantly different from zero. OPMD- and age-associated changes were defined as those with δ significantly different from zero. To determine whether the expression profiles of individual genes significantly differ between controls and OPMD *P*-values are FDR-corrected with the cut-off threshold of 0.05.

Quantitative RT-PCR analysis

Primers for Q-PCR validation were designed in the sequence surrounding the Illumina probe location using Primer 3 plus program. RT-QPCR was performed according to the procedure in Trollet et al. (2010). The list of primers is provided in Table 3.

Cell culture and transfection

C2C12 cells were used for transient transfection experiments. C2C12 cells were cultured in DMEM containing 20% fetal calf serum. Prior to transfection, cells were seeded on glass. Transfection was carried out in 80% cell confluence with Lipofectamine$^{\text{TM}}$ 2000 (Invitrogen) according to the manufacturer's protocol. Plasmids used for transfection are YFP-Ala16-PABPN1 and Histone4-CFP. For the proteasome modification treatments, cells were treated 16 hours after transfection with DMSO (1:1000), 5 μM MG132 (Sigma-Aldrich), or 5U/ml IFNγ (HyCult Biotech) for 20 hours.

Protein detection and Imaging

For immunocytochemistry, 16 hours post-transfection with YFP-Ala16-PABPN1, C2C12 cells were incubated with fusion medium (DMEM supplemented with 2.5% Horse serum) for 2 days, and immunocytochemistry was performed after a short fixation (Raz et al., 2006) followed by a

15 min incubation with 1% Triton X100, during which PABPN1 aggregates remain intact. Following antibody incubations, preparations were mounted in Citifluor (Agar Scientific) containing 400 μg/ml of DAPI (Sigma-Aldrich). Immunofluorescent specimens were examined with a fluorescence microscope (Leica DM RXA), 63X and 100X lens NA 1.4 plan Apo objective. Integrated intensity was measured with ImageJ (http://rsbweb.nih.gov/ij/), and intensity values were corrected for background.

Antibodies used in this study are: Goat anti-Asb-11 (K16) (1:1000) Santa Cruz Biotechnology; Rabbit anti-atrogin-1 (1:1000) ECM Biosciences; Rabbit anti-Murf1 (1:1000) ECM Biosciences; Goat anti-DDB1 (1:1000) Abcam; mouse anti-Flag (1:2000) Sigma –Aldrich; Rabbit anti-Desmin MP Biomedicals. Alexa-Fluor 594 conjugated secondary (Invitrogen) or IRDye 680LT and 800CW conjugated secondaries (Licor Biosciences) were used to detection of first antibody.

Table 3 - The list of primers used for quantitative RT_PCR analysis.

Genes	Probe	FW Primer Sequence	RV Primer Sequence
Arih1	6900025	GAGAAGGATGGCGGTTGTAA	ATCTCTTGCTGCCTTTGCAT
Arih2	2810025	AGCCTAACTCCCCCTTGGTA	ACCACTGAGGGTGCAAAAAC
Ate1	6940722	CAAAGTGATTCTACTGTGGCTGA	ACGAAAATCTCCAATGCAGTC
Cul7	3360114	CGGGACTATGCGGTGATACT	GTGGGTTCGTCTGTGGTCTT
Psme3	2810537	GCGAAGGTCAAACCCATAGA	GAAAGTGATGCATCCCAGGT
Rbx1	2340047	TTGAGGCCAGCCTACAGAGT	AGGAAAACTCCCCTGAAGGA
Skp1a	2450102	TGCAGCTGGGCTCTCTTAAT	GTTTCTCCACCTGGGAACAA
Uchl1	1230066	CCTGTCCCTTCAGTTCCTCA	GATTAACCCCGAGATGCTGA
Huwe1	106840041	GCTGCATTGAGACTTGAAACC	TCCACAACACAGATGCCAAT
Tbl1x	6400524	ATTTTCCCCCTCCCCTAATC	GAGCCTGTTCTGGATGGAAA
Ube4b	3610154	GCTGGAGTGGATCAGGACTC	TGGTAAGGTCAAACCCCAAA
Ube2o	2190040	CGGTGAGCACATTACAGCTC	GCATCATGCTTTGGCTTTTT
Usp47	100940601	GAATGCTTGTAAAGTCCCGTTT	CTAGCACGCTCTGCAATGAA
Ppp2cb	5570593	ACTGCTACCGTTGTGGGAAC	AGGTCCTGGGGAGGAATTTA
Ube3b	6380458	GCCTGCACAGGTAACACAGA	ACCAGGAGCTGCTGAGATGT
Fbxo32	110037	GGGAGGCAATGTCTGTGTTT	AAGAGGTGCAGGGACTGAGA
Trim63	1740164	CGACCGAGTGCAGACGATCATCTC	GTGTCAAACTTCTGACTCAGC
Ubr5	1780605	GCTGCCTTTGTGGAAAGTGT	TTGCAGCCAACCACAAATAA
Asb11	2060487	TTGTGCTGAACAAGCTCCTG	GAGGGTCCTGAATCATCCAA
mHPRT	-	CGTCGTGATTAGCGATGATG	TTTTCCAAATCCTCGGCATA

Authors' contributions

SYA and AV preformed the bioinformatics studies. AV and VR preformed the molecular genetics studies. Biological samples were provided by: BS, BE, MS, JV, CT, GD, AC and MS. The manuscript was drafted by SYA and VR and written by SYA, PAC, SM and VR. PAC participated in the bioinformatics design, coordination and data analysis. All authors read and approved the manuscript.

Acknowledgement

This work was supported by grants from European Commission (PolyALA LSHM-CT-2005018675) and Muscular Dystrophy Association (68016) to S.M.M. and MDA, the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO), the CNRS (UPR1142), the ANR Genopat (ANR-09-GENO-025-01), the FRM ("Equipe FRM 2007" N°DEQ20071210560), and the European Commission (PolyALA LSHM-CT-2005-018675) to M.S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reference List

Abu-Baker,A., Messaed,C., Laganiere,J., Gaspar,C., Brais,B., and Rouleau,G.A. (2003). Involvement of the ubiquitin-proteasome pathway and molecular chaperones in oculopharyngeal muscular dystrophy. Hum. Mol. Genet *12*, 2609-2623.

Blumen,S.C., Bouchard,J.P., Brais,B., Carasso,R.L., Paleacu,D., Drory,V.E., Chantal,S., Blumen,N., and Braverman,I. (2009). Cognitive impairment and reduced life span of oculopharyngeal muscular dystrophy homozygotes. Neurology *73*, 596-601.

Bodine,S.C., Latres,E., Baumhueter,S., Lai,V.K., Nunez,L., Clarke,B.A., Poueymirou,W.T., Panaro,F.J., Na,E., Dharmarajan,K., Pan,Z.Q., Valenzuela,D.M., DeChiara,T.M., Stitt,T.N., Yancopoulos,G.D., and Glass,D.J. (2001). Identification of ubiquitin ligases required for skeletal muscle atrophy. Science *294*, 1704-1708.

Brais,B., Bouchard,J.P., Xie,Y.G., Rochefort,D.L., Chretien,N., Tome,F.M., Lafreniere,R.G., Rommens,J.M., Uyama,E., Nohira,O., Blumen,S., Korczyn,A.D., Heutink,P., Mathieu,J., Duranceau,A., Codere,F., Fardeau,M., and Rouleau,G.A. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet *18*, 164-167.

Brais,B., Xie,Y.G., Sanson,M., Morgan,K., Weissenbach,J., Korczyn,A.D., Blumen,S.C., Fardeau,M., Tome,F.M., Bouchard,J.P., and . (1995). The oculopharyngeal muscular dystrophy locus maps to the region of the cardiac alpha and beta myosin heavy chain genes on chromosome 14q11.2-q13. Hum. Mol. Genet *4*, 429-434.

Calado,A., Tome,F.M., Brais,B., Rouleau,G.A., Kuhn,U., Wahle,E., and Carmo-Fonseca,M. (2000). Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. Hum. Mol. Genet *9*, 2321-2328.

Cao,P.R., Kim,H.J., and Lecker,S.H. (2005). Ubiquitin-protein ligases in muscle wasting. Int. J. Biochem. Cell Biol. *37*, 2088-2097.

Catoire,H., Pasco,M.Y., Abu-Baker,A., Holbert,S., Tourette,C., Brais,B., Rouleau,G.A., Parker,J.A., and Neri,C. (2008). Sirtuin inhibition protects from the polyalanine muscular dystrophy protein PABPN1. Hum. Mol. Genet *17*, 2108-2117.

Chartier,A., Benoit,B., and Simonelig,M. (2006). A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. EMBO J *25*, 2253-2262.

Chartier,A., Raz,V., Sterrenburg,E., Verrips,C.T., van der Maarel,S.M., and Simonelig,M. (2009). Prevention of oculopharyngeal muscular dystrophy by muscular expression of Llama single-chain intrabodies in vivo. Hum. Mol. Genet.

Ciechanover,A. and Brundin,P. (2003). The ubiquitin proteasome system in neurodegenerative diseases: sometimes the chicken, sometimes the egg. Neuron *40*, 427-446.

Combaret,L., Dardevet,D., Bechet,D., Taillandier,D., Mosoni,L., and Attaix,D. (2009). Skeletal muscle proteolysis in aging. Curr. Opin. Clin. Nutr. Metab Care *12*, 37-41.

Corbeil-Girard,L.P., Klein,A.F., Sasseville,A.M., Lavoie,H., Dicaire,M.J., Saint-Denis,A., Page,M., Duranceau,A., Codere,F., Bouchard,J.P., Karpati,G., Rouleau,G.A., Massie,B., Langelier,Y., and Brais,B. (2005). PABPN1 overexpression leads to upregulation of genes encoding nuclear proteins that are sequestered in oculopharyngeal muscular dystrophy nuclear inclusions. Neurobiol. Dis. *18*, 551-567.

Davies,J.E., Sarkar,S., and Rubinsztein,D.C. (2006). Trehalose reduces aggregate formation and delays pathology in a transgenic mouse model of oculopharyngeal muscular dystrophy. Hum. Mol. Genet *15*, 23-31.

Davies,J.E., Wang,L., Garcia-Oroz,L., Cook,L.J., Vacher,C., O'Donovan,D.G., and Rubinsztein,D.C. (2005). Doxycycline attenuates and delays toxicity of the oculopharyngeal muscular dystrophy mutation in transgenic mice. Nat Med. *11*, 672-677.

Dennis,G., Jr., Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C., and Lempicki,R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, 3.

Fan,X. and Rouleau,G.A. (2003). Progress in understanding the pathogenesis of oculopharyngeal muscular dystrophy. Can. J. Neurol. Sci. *30*, 8-14.

Ferrington,D.A., Husom,A.D., and Thompson,L.V. (2005). Altered proteasome structure, function, and oxidation in aged muscle. FASEB J. *19*, 644-646.

Finley,D. (2009). Recognition and processing of ubiquitin-protein conjugates by the proteasome. Annu. Rev Biochem. *78*, 477-513.

Goeman,J.J., van de Geer,S.A., de,K.F., and van Houwelingen,H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. *20*, 93-99.

Hino,H., Araki,K., Uyama,E., Takeya,M., Araki,M., Yoshinobu,K., Miike,K., Kawazoe,Y., Maeda,Y., Uchino,M., and Yamamura,K. (2004). Myopathy phenotype in transgenic mice expressing mutated PABPN1 as a model of oculopharyngeal muscular dystrophy. Hum. Mol. Genet *13*, 181-190.

Hoeller,D. and Dikic,I. (2009). Targeting the ubiquitin system in cancer therapy. Nature *458*, 438-444.

Huang,d.W., Sherman,B.T., and Lempicki,R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. *4*, 44-57.

Ioannidis,J.P., Allison,D.B., Ball,C.A., Coulibaly,I., Cui,X., Culhane,A.C., Falchi,M., Furlanello,C., Game,L., Jurman,G., Mangion,J., Mehta,T., Nitzberg,M., Page,G.P., Petretto,E., and van,N., V (2009). Repeatability of published microarray gene expression analyses. Nat Genet. *41*, 149-155.

Jelier,R., Schuemie,M.J., Veldhoven,A., Dorssers,L.C., Jenster,G., and Kors,J.A. (2008). Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. *9*, R96.

Kloetzel,P.M. and Ossendorp,F. (2004). Proteasome and peptidase function in MHC-class-I-mediated antigen presentation. Curr. Opin. Immunol. *16*, 76-81.

Lee,C.K., Klopp,R.G., Weindruch,R., and Prolla,T.A. (1999). Gene expression profile of aging and its retardation by caloric restriction. Science *285*, 1390-1393.

Li,W., Bengtson,M.H., Ulbrich,A., Matsuda,A., Reddy,V.A., Orth,A., Chanda,S.K., Batalov,S., and Joazeiro,C.A. (2008). Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling. PLoS. One. *3*, e1487.

Liu,Z., Miers,W.R., Wei,L., and Barrett,E.J. (2000). The ubiquitin-proteasome proteolytic pathway in heart vs skeletal muscle: effects of acute diabetes. Biochem. Biophys. Res. Commun. *276*, 1255-1260.

Osna,N.A., Clemens,D.L., and Donohue,T.M., Jr. (2003). Interferon gamma enhances proteasome activity in recombinant Hep G2 cells that express cytochrome P4502E1: modulation by ethanol. Biochem. Pharmacol. *66*, 697-710.

Raz,V., Carlotti,F., Vermolen,B.J., van der,P.E., Sloos,W.C., Knaan-Shanzer,S., de Vries,A.A., Hoeben,R.C., Young,I.T., Tanke,H.J., Garini,Y., and Dirks,R.W. (2006). Changes in lamina structure are followed by spatial reorganization of heterochromatic regions in caspase-8-activated human mesenchymal stem cells. J Cell Sci. *119*, 4247-4256.

Reyes-Turcu,F.E., Ventii,K.H., and Wilkinson,K.D. (2009). Regulation and cellular roles of ubiquitin-specific deubiquitinating enzymes. Annu. Rev Biochem. *78*, 363-397.

Sandri,M. (2008). Signaling in muscle atrophy and hypertrophy. Physiology. (Bethesda.) *23*, 160-170.

Smyth,G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Smyth,G.K. and Speed,T. (2003). Normalization of cDNA microarray data. Methods *31*, 265-273.

Taillandier,D., Combaret,L., Pouch,M.N., Samuels,S.E., Bechet,D., and Attaix,D. (2004). The role of ubiquitin-proteasome-dependent proteolysis in the remodelling of skeletal muscle. Proc. Nutr. Soc. *63*, 357-361.

Tavanez,J.P., Calado,P., Braga,J., Lafarga,M., and Carmo-Fonseca,M. (2005). In vivo aggregation properties of the nuclear poly(A)-binding protein PABPN1. RNA. *11*, 752-762.

Tome,F.M. and Fardeau,M. (1980). Nuclear inclusions in oculopharyngeal dystrophy. Acta Neuropathol. *49*, 85-87.

Trollet,C., Anvar,S.Y., Venema,A., Hargreaves,I.P., Foster,K., Vignaud,A., Ferry,A., Negroni,E., Hourde,C., Baraibar,M.A., 't Hoen,P.A., Davies,J.E., Rubinsztein,D.C., Heales,S.J., Mouly,V., van der Maarel,S.M., Butler-Browne,G., Raz,V., and Dickson,G. (2010). Molecular and phenotypic characterization of a mouse model of oculopharyngeal muscular dystrophy reveals severe muscular atrophy restricted to fast glycolytic fibres. Hum. Mol. Genet.

VICTOR,M., HAYES,R., and ADAMS,R.D. (1962). Oculopharyngeal muscular dystrophy. A familial disease of late life characterized by dysphagia and progressive ptosis of the evelids. N. Engl. J. Med. *267*, 1267-1272.

Vignaud,A., Ferry,A., Huguet,A., Baraibar,M., Trollet,C., Hyzewicz,J., Butler-Browne,G., Puymirat,J., Gourdon,G., and Furling,D. (2010). Progressive skeletal muscle weakness in transgenic mice expressing CTG expansions is associated with the activation of the ubiquitin-proteasome pathway. Neuromuscul. Disord. *20*, 319-325.

APPENDIX

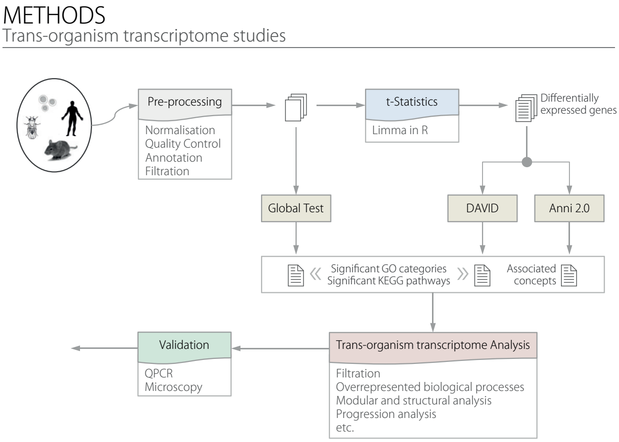
Supplementary Table 1A - Overview of genome-wide transcriptome microarray datasets of *Drosophila* and mouse OPMD models and muscle biopsies of OPMD patients.

Biological Systems	Tissue	Number of Samples		Age	MA Platform
Drosophila	Adult thoracic muscles	3	pools of 50 flies per genotype	1 day	15K INDAC spotted oligonucleotide array
				6 days	
				11 days	
Mouse	Quadriceps	6	replicates per genotype	6 weeks	Illumina 48K Mouse v.1 bead array
				18 weeks	
				26 weeks	
Human	Quadriceps	9	Pre-symptomatic	17 – 22 years control	Illumina 48K Human v.3 bead array
		13	Symptomatic	31 – 40 years Pre-Symptomatic	
		39	Controls	38 – 42 years control	
				49 – 60 years symptomatic	
				58 – 67 years control	

Supplementary Table 1B - Overview of muscle biopsies of OPMD patients and controls. All patients are heterozygous expPABPN1 carriers as indicated by sequence analysis.

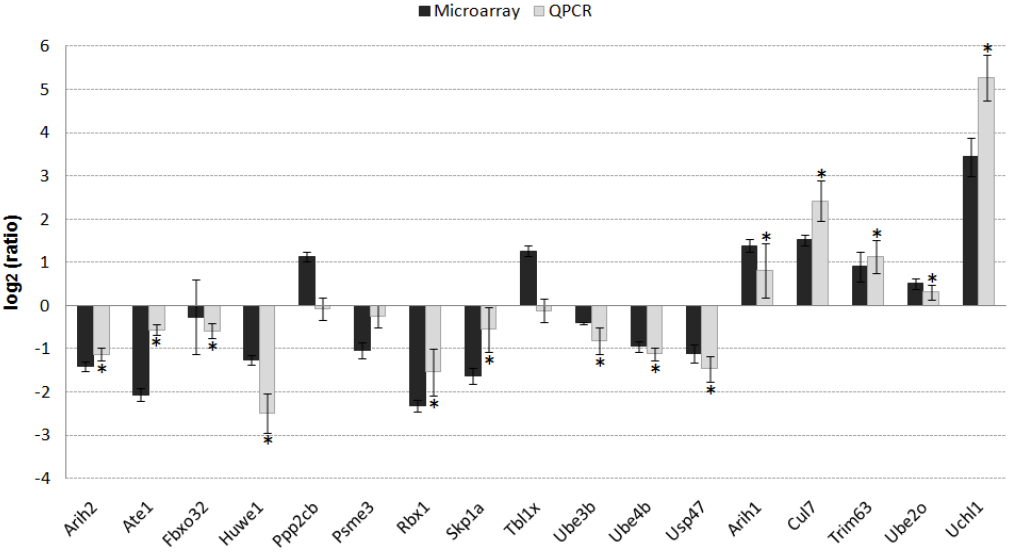
	Sex	Age	GCG Mutation	Muscle Histology
Pre-Symptomatic	Female	39	12/6	Sporadic atrophic fibre
	Female	37	10/6	Moderate dystrophic alterations
	Female	37	12/6	Normal
	Female	31	9/6	Slight dystrophic alteration
Symptomatic	Female	60	9/6	Moderate dystrophic alterations
	Female	49	10/6	Moderate dystrophic alterations
	Male	59	10/6	Moderate dystrophic alterations
	Female	57	11/6	Severe dystrophic alterations

Supplementary Figure 1 - Integrated cross-species high-throughput transcriptome study.


METHODS
Trans-organism transcriptome studies
Pre-processing
Normalisation
Quality Control
Annotation
Filtration
t-Statistics
Limma in R
Differentially expressed genes
Global Test
DAVID
Anni 2.0
Significant GO categories
Significant KEGG pathways
Associated concepts
Validation
QPCR
Microscopy
Trans-organism transcriptome Analysis
Filtration
Overrepresented biological processes
Modular and structural analysis
Progression analysis
etc.

Supplementary Table 2 - Literature-aided analysis of the association of biomedical concepts with OPMD-deregulated genes.

	Drosophila		Mouse		Human
1	ribosomal protein activity	1	Ubiquitination	1	ubiquitin activity
2	ubiquitin activity	2	Ubiquitins	2	RNA Binding
3	RNA Binding	3	Ubiquitin	3	Ubiquitination
4	Ubiquitination	4	Ligase	4	RNA Splicing
5	polyethylene glycol monostearate	5	SNAP receptor	5	GTP Binding
6	Ligase	6	Internal Ribosome Entry Site	6	protein transport
7	POLR2F	7	Phosphotransferases	7	Alternative Splicing
8	GTP Binding	8	Cullin Proteins	8	Transcription, Genetic
9	ribosome biogenesis and assembly	9	Muscle Proteins	9	intracellular protein transport
10	Ribosome Subunits	10	Mitogen-Activated Protein Kinases	10	GTP-binding



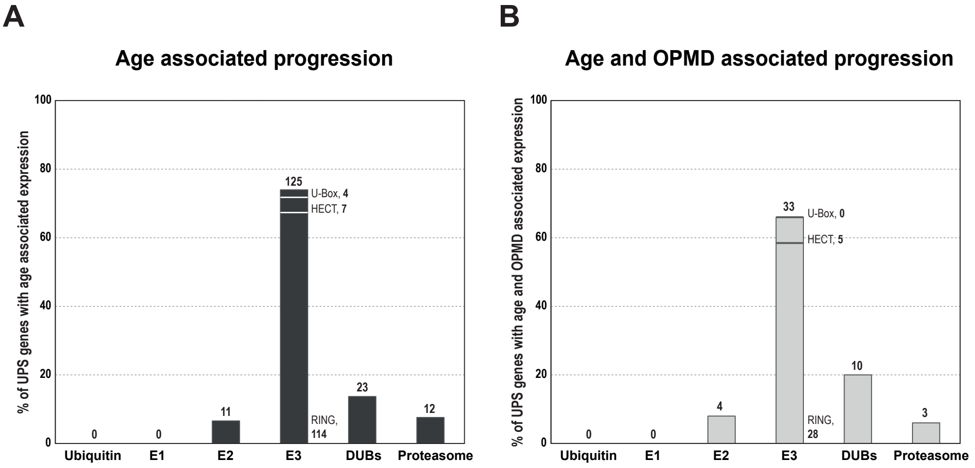
Supplementary Figure 2 - Validation of expression level of selected genes from the pool of UPS OPMD-deregulated genes on the skeletal muscle of 6 weeks-old OPMD mice, normalized to WT. Histograms indicate the log2(ratio) of the measured expression values using RT Q-PCR (grey bars) and microarray (black bars) for 4 WT and 6 OPMC mice (∗ $P < 0.05$).

Supplementary Table 3 - Spreading of OPMD-deregulation in UPS over the functional components and sub-classes of E3-ligases. The number of total genes in each component is shown; the percentage of OPMD-deregulated (D.E.) genes and *P*-values are indicated. For *Drosophila* and mouse, *P*-values indicate the significance of OPMD-deregulation in combined datasets from three time-points and the percentage of D.E. genes represents an average across three time-points. The overlap in OPMD-deregulated genes between human and mouse is shown and the percentage of deregulated genes shows the fraction of overlap in human D.E. genes.

										Overlap	
	Drosophila			**Mouse**			**Human**			**Mouse vs. Human**	
	# Total Genes	% D.E. Genes	*P*-Value (FDR)	# Total Genes	% D.E. Genes	*P*-Value (FDR)	# Total Genes	% D.E. Genes	*P*-Value (FDR)	# D.E. Genes	% D.E. Genes
Ubiquitin	2	50.00	4.18E-05	3	11.11	1.28E-01	3	33.33	1.19E-01	0	00.00
E1 Ubiquitin Activation	4	16.67	1.24E-01	7	04.76	7.29E-02	7	14.29	7.94E-02	0	00.00
E2 Ubiquitin Conjugation	22	13.64	9.19E-06	33	44.42	1.51E-08	34	23.53	7.31E-02	3	37.50
E3 Ubiquitin Ligase	249	13.99	1.92E-04	526	29.58	1.64E-08	586	24.74	4.35E-03	69	47.59
RING	226	13.61	1.29E-04	495	28.13	1.67E-08	550	23.04	7.83E-03	60	47.35
HECT	16	22.92	2.49E-02	23	42.00	1.68E-08	28	32.14	1.59E-02	7	77.78
U-Box	7	4.76	6.98E-01	8	46.00	2.52E-08	8	37.50	7.94E-03	2	66.67
Deubiquitination (DUB)	35	20.00	1.63E-05	72	45.83	1.48E-08	75	24.00	3.15E-02	13	72.22
Proteasome	47	29.79	2.15E-04	37	36.94	1.37E-07	37	51.35	9.27E-03	11	57.90

Supplementary Table 4 - Direction of OPMD-deregulation over the functional components of UPS that are significantly deregulated in all organisms.

	Mouse				Human			
	# Total Genes	% D.E. Genes	UP	DOWN	# Total Genes	% D.E. Genes	UP	DOWN
E2 Ubiquitin Conjugation	33	44.42	50.00	50.00	34	23.53	50.00	50.00
E3 Ubiquitin Ligase	526	29.58	51.44	48.56	586	24.74	50.00	50.00
Deubiquitination (DUB)	72	45.83	42.00	58.00	75	24.00	41.18	58.82
Proteasome	37	36.94	59.38	40.62	37	51.35	42.86	57.14



Supplementary Figure 3: Temporal changes in UPS gene expression. A linear regression was applied to identify temporal changes in expression levels of OPMD-deregulated genes in A17.1. **A)** Histogram shows the percentage of age associated OPMD-deregulated genes for each of the UPS functional components and E3-ligase subclasses. **B)** OPMD-deregulated genes showing age and OPMD associated progression. Histogram shows the percentage of genes with age and OPMD associated expression for each of the functional components. The number of genes in each bar is indicated.

Supplementary Table 5 - Deregulation of autophagy and lysosome as compared to proteasome in OPMD model systems and OPMD patients.

									Overlap	
	Drosophila			Mouse			Human		Mouse vs. Human	
	# Total Genes	% D.E. Genes	P-Value (FDR)	# Total Genes	% D.E. Genes	P-Value (FDR)	# Total Genes	% D.E. Genes	P-Value (FDR)	# D.E. Genes	% D.E. Genes
Proteasome	47	29.79	2.15E-04	37	36.94	1.37E-07	37	51.35	9.27E-03	11	57.90
Autophagy	16	25.00	1.07E-03	39	30.77	8.13E-08	32	18.75	1.37E-02	1	16.67
Lysosome	60	5.00	1.64E-02	75	25.33	6.06E-03	73	24.66	1.54E-02	6	33.33

Additional files

Title: A list of deregulated UPS genes in mouse and human datasets.

Description: This additional file contains t-statistics (fold change and P value) for genes within the ubiquitin-proteasome pathway, in human and mouse datasets. The t-statistics represent the association of the gene expression profiles to OPMD.

Link: *http://www.skeletalmusclejournal.com/content/supplementary/2044-5040-1-15-s2.xlsx*

Decline in PABPN1 expression level marks skeletal muscle aging

Seyed Yahya Anvar[1,*], Yotam Raz[2], Andrea Venema[1], Merel L.R. van 't Hoff[1], Marius Gheorghe[1], Jelle J. Goeman[3], Barbara van der Sluijs[4], Baziel van Engelen[4], Marc Snoeck[5], John Vissing[6], Silvère M. van der Maarel[1], Peter A.C. 't Hoen[1] and Vered Raz[1,*]

Aging-associated disorders can be accompanied by increased tissue degeneration and may provide insight into key regulators of aging. Oculopharyngeal muscular dystrophy (OPMD) is caused by alanine-expansion mutations in *PABPN1,* and is characterized by progressive skeletal muscle weakness that is manifested after midlife. We compared expression profiles from *Vastus lateralis* of controls and OPMD. Similar to PABPN1 expression, between 40-45 years a transcriptional switch was identified in both OPMD and muscle aging while trends in OPMD were accelerated. Among these genes, we identified a significant and progressive decline in *PABPN1* expression from the fifth decade in aging muscles. In concurrence with the more severe muscle weakness, this decline was accelerated in muscles primarily affected in OPMD. The aging-associated decline of *PABPN1* was not detected in other tissues or in blood from OPMD patients. We show that down-regulation of *PABPN1* induced progressive cell senescence in myoblast cultures. We suggest that a decline in *PABPN1* expression marks muscle aging and reduced levels of the protein causes age-associated muscle degeneration.

---

1 Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. **2** Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, the Netherlands. **3** Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands. **4** Department of Neurology, Radboud University Nijmegen Medical Center, Nijmegen, the Netherlands. **5** Department of Anaesthesia, Canisius-Wilhelmina Hospital, Nijmegen, the Netherlands. **6** Neuromuscular Research Unit and Department of Neurology, Rigshospitalet Copenhagen, Copenhagen, Denmark.

* To whom correspondence should be addressed at: v.raz@lumc.nl

Manuscript Submitted, 2012

INTRODUCTION

Aging is marked by a progressive decline of cellular activities and its rate differs between tissues (Kirkwood and Austad, 2000). A decrease in skeletal muscle performance, as measured by strength, highly correlates with biological aging. Age-associated muscle weakness in healthy cohorts starts around the fifth decade and linearly progresses with age (Beenakker et al., 2010). A decline in muscle strength is suggested to predict functional disability and mortality in elderly (Liu and Latham, 2011; Ling et al., 2010; Roth et al., 2002). The degenerative loss of muscle function during aging is regulated by numerous genetic and environmental factors. Consequently, the onset and progression of aging-associated decline in muscle performance vary greatly between individuals. Aging is a complex process and the molecular mechanisms that control the onset and progression of muscle aging, as well as key regulators, are not fully understood. The high complexity of aging-associated molecular mechanisms is demonstrated by genome-wide changes in mRNA expression affecting a broad range of biological processes. Genome-wide transcriptional changes can be derived by changes in mRNA stability. Thus, it is expected that regulators of mRNA processing would regulate aging-associated transcriptional changes.

Aging associated changes can sometimes be exacerbated in patients with late onset degenerative disorders (Kirkwood and Austad, 2000). Studies of late onset disorders can thereby expose key regulators of aging that are otherwise difficult to identify. Oculopharyngeal muscular dystrophy (OPMD) is a late onset autosomal dominant muscle disorder. OPMD is characterized by progressive ptosis, dysphagia, and proximal limb muscle weakness that typically appear from the fifth decade (Brais et al., 1995; Taylor, 1915; van der Sluijs et al., 2003). OPMD is caused by a trinucleotide repeat expansion mutation in the gene encoding for *Poly(A) Binding Protein Nuclear 1* (*PABPN1*) causing a poly-alanine expansion in the N-terminus of PABPN1 (expPABPN1) (Brais et al., 1998). PABPN1 binds to mRNA and regulates poly(A) elongation (Benoit et al., 2005). The length of poly(A) depends on PABPN1 concentration (Kuhn et al., 2009), and knockdown of PABPN1 causes shortening of poly(A) tail mRNA (Apponi et al., 2010). PABPN1 knockdown in mouse myotubes leads to myogenic defects and reduced cell fusion (Apponi et al., 2010). Reduced cell fusion was also reported in OPMD myoblast cultures (Perie et al., 2006). Overexpression of mutant PABPN1 also leads to muscle cell defects in a mouse model (Davies et al., 2005; Trollet et al., 2010). Mutant PABPN1 is prone to aggregation and accumulates in insoluble nuclear inclusions (Tome and Fardeau, 1980). Although prevention of protein aggregation in animal models with high overexpression of expPABPN1 are effective in delay of muscle weakness (Davies et al., 2005; Chartier et al., 2009; Catoire et al., 2008), aggregation of wild-type PABPN1 were also reported in aging rat neuron cells (Berciano et al., 2004). In contrast to aggregates of expAPBPN1, those of the wild type protein are not disease-associated. In cell models both wild type and expPABPN1 form aggregates, while expPABPN1 is more prone to aggregation (Raz et al., 2011a; Raz et al., 2011b). Differences in aggregation can be, in part, explained in differences in poly-ubiquitination (Raz et al., 2011b). Inhibition of the proteasome enhances the aggregation of expPABPN1 in cell models (Abu-Baker et al., 2003; Raz et al., 2011b). In OPMD the ubiquitin-proteasome system (UPS) is significantly deregulated (Anvar et al., 2011; Raz et al., 2011b). Dysfunctional UPS stimulates the formation of many protein aggregates (Balch et al., 2008; Morimoto, 2008; Sherman and Goldberg, 2001).

It is unclear how a ubiquitously expressed protein, like PABPBN1, predominantly affects only a subset of skeletal muscles and causes symptoms that are not apparent until midlife. We hypothesized that aging contributes to the initiation and progressiveness of muscle weakness in OPMD. We investigated the hypothesis that aging factors contribute to OPMD. We identified significant similarities between OPMD-deregulated and aging–regulated expression profiles. In concurrence with muscle symptoms in OPMD, transcriptional changes were accelerated in OPMD compared with normal aging. We show that a decline in PABPN1 expression is highly correlated with age-associated changes in muscle strength in both OPMD and in muscle aging. We show that down regulation of PABPN1 induces cell senescence. Since PABPN1 regulates mRNA stability, we suggest that changes in PABPN1 expression levels in muscle cells would lead to broad transcriptional changes and hence muscle weakness.

RESULTS

Molecular signatures of aging are found in the OPMD mouse model at young age

Symptoms in OPMD do not become apparent until midlife. Therefore, we hypothesised that molecular processes that control muscle aging are involved in OPMD pathogenesis. We investigated whether aging-regulated genes are deregulated in a mouse model for OPMD. In the A17.1 mouse expPABPN1 is overexpressed in muscles leading to muscle weakness (Davies et al., 2005). In this mouse model, muscle atrophy initiates after 12 weeks (Trollet et al., 2010). A17.1-deregulated genes were identified from age-matched wild type controls (Trollet et al., 2010). In a literature-aided association study (LAS), we observed a large subset of A17.1-deregulated genes, in 6 week-old A17.1 mice, that were strongly associated with the term '*Aging*' (**Figure 1A**). Moreover, the fold-change of these genes was remarkably high (**Figure 1A**). This suggests that in this mouse model aging-associated transcriptional changes are induced already at 6 weeks. In an unsupervised meta-analysis, 104 microarray studies, which are related to muscle development and muscle disorders, were compared with that of A17.1. Three major clusters of similar transcriptional changes were identified (**Figure 1B**). The transcriptome of the 6 week-old A17.1 mouse was clustered together with those related to skeletal muscle aging (Welle et al., 2004; Giresi et al., 2005), but not with datasets from other muscular dystrophies or myopathies (**Figure 1B**). These analyses further indicate that transcriptional changes in OPMD are highly associated with those of muscle aging.

Common molecular signatures in muscle aging and OPMD

To investigate genome-wide transcriptional changes in OPMD and during aging in humans, three microarray datasets were generated from *Vastus Lateralis* muscles. For muscle aging a continuous cross sectional dataset was generated from controls aged 17-89. Datasets from OPMD and expPABPN1 carries at pre-symptomatic stage were generated after comparing to age-matching control groups (**Supplementary Table 1**). Major sources of transcriptional variation were assessed using unsupervised principal component analysis (PCA). In the control dataset age-associated variations were identified using the first three principal components, covering 49% of transcriptional variation. Based on the PCA analysis, samples were clustered into two age groups of 17-42 and 43-89 years (**Figure 2A**). This suggests a genome-wide transcriptional switch at the first half of the fifth decade. To verify this, we analysed the expression trends of probes whose expression changed with age (named here as aging-regulated; $P<0.05$). We identified a major switch-point around the age of 42±5 years (**Figure 2A**). An absolute correlation distance measure of k-means clustering revealed that the up-regulated and down-regulated trends of 70% of the age-regulated probes are crossed at 42±5 years (**Figure 2B**). This indicates that a major expression switch in skeletal muscles occurs during the first half of the fifth decade. This observation is in agreement with physiological studies in continuous cross-sectional cohorts showing that aging-related changes in muscle strength start between 40 to 50 years (Kirkwood, 2005; Lexell et al., 1988; Lindle et al., 1997; Sahin and Depinho, 2010). The aging-regulated genes were mapped to a wide spectrum of Kyoto Encyclopaedia of Genes and Genomes (KEGG) functional pathways.

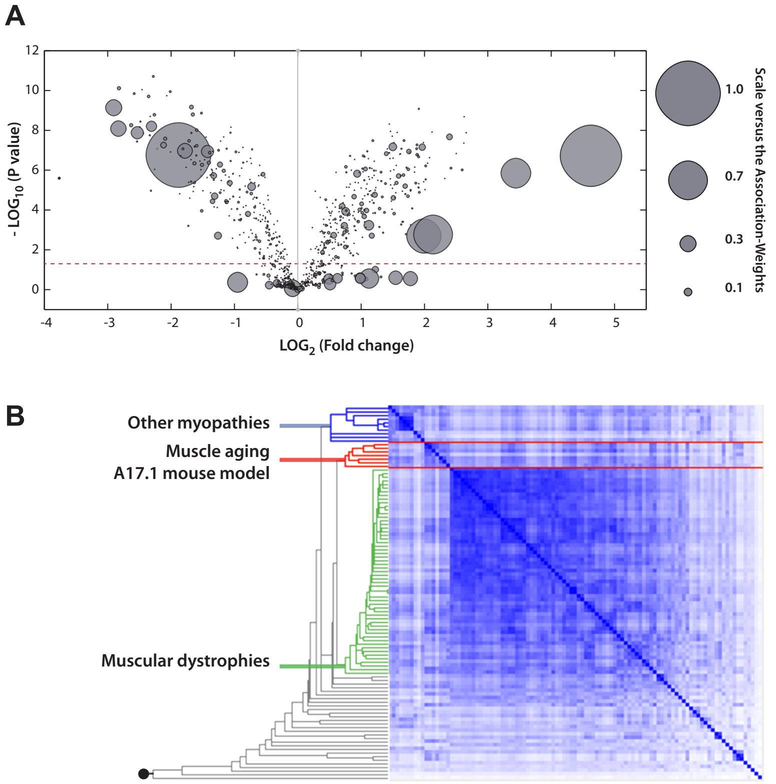


Figure 1 - The A17.1 mouse transcriptome is strongly associated with aging. A) Volcano plot shows the distribution of significantly deregulated genes ($P = 0.05$; indicated with a dashed line) in 6 week-old A17.1 mice against fold change. Genes are weighted based on their association with the *Aging* concept. The normalized association-weight is presented with a circle on a scale between 0.05 and 1, where 1 equals the highest association. **B)** Hierarchical clustering arrangements of 104 datasets in a literature-aided meta-analysis. Shades of blue indicate degree of similarities: from weak (white) to strong (dark blue). Three skeletal muscle aging-related datasets are clustered with OPMD dataset of 6 week-old mice (highlighted in red). The clusters associated with muscular dystrophies and other myopathies are highlighted in green and blue, respectively.

These aging-regulated KEGG pathways were highly similar to those that were identified from independent microarray study of skeletal muscles from two-age group (Welle et al., 2004; Welle et al., 2003; **Supplementary Table 2**).

Around midlife, muscle weakness symptoms are found in OPMD but not in age-matching controls (van der Sluijs et al., 2003) or in expPABPN1 carriers at a pre-symptomatic stage (**Supplementary Table 2**). OPMD-deregulated or pre-symptomatic-deregulated genes were identified from age-matching controls. Despite the limited number of samples in OPMD, OPMD-deregulated genes were highly similar to those identified in OPMD animal models (Anvar et al., 2011; Raz et al., 2011b). In OPMD large transcriptional changes were identified, but only minor transcriptional changes was identified at the pre-symtomatic stage (**Figure 2C**). Only 9% of the OPMD-deregulated genes were also deregulated in the pre-symptomatic (**Figure 2C**). 30 KEGG pathways were enriched in OPMD-deregulated genes (**Supplementary Table 2**), whereas no con-

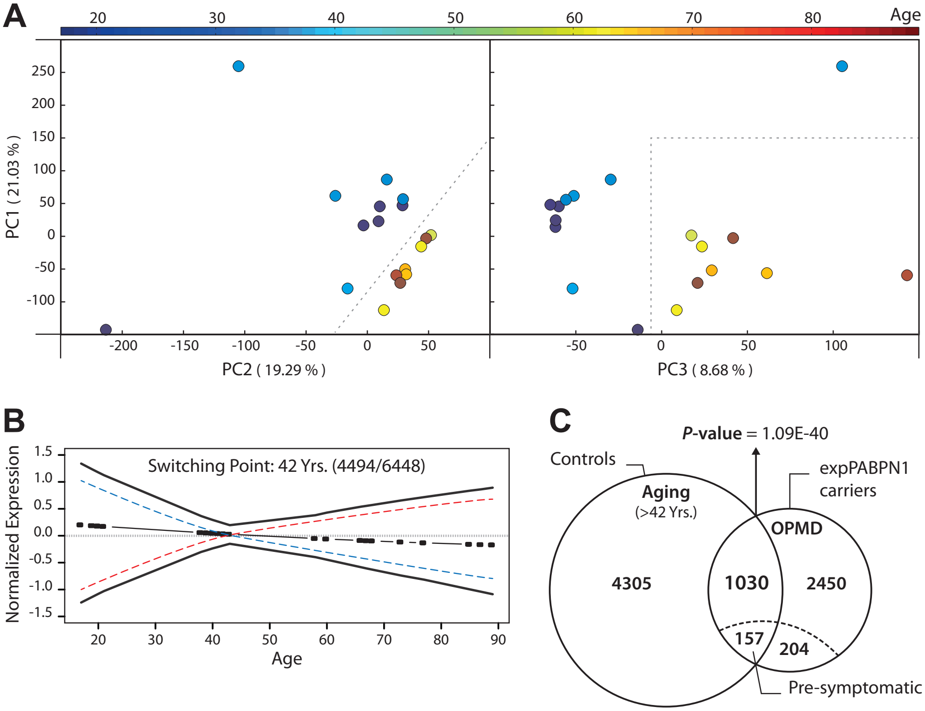


Figure 2 - High similarities between transcriptomes of muscle aging and OPMD. A) Principal component analysis (PCA) plots of skeletal muscle datasets from healthy controls (age is indicated with a colour scale). An age-associated variation is found with the first three principal components. Plots show sample distribution in the first and second (left) or first and third (right) components. The percentage of variations is indicated between brackets. The colour scale reflecting the age of the patient samples is given on top of the figure. Dashed lines separate samples into two age groups. **B)** Plot shows expression trends for the major cluster of 6448 probes whose expression are significantly changed with age (*P*<0.05). 4494 probes whose expression significantly change with age (p<0.05) were used for k-mean clustering analysis. Similar trends with up- and or down-regulation were combined using absolute correlation, revealing a switching point at 42±5 years. Up- or down- regulated expression trends (red and blue, respectively) are indicated with dashed lines, and continuous lines show the 95% boundaries. The middle line indicates the centroid with the age of individual samples. **C)** Venn diagram shows the overlap of between genes associated with aging (>42) and differentially expressed genes between OPMD- or expPABPN1 carriers and age-matched controls. Differentially expressed genes (*P*<0.05) in OPMD and pre-symptomatic carriers were identified from age matching control groups. P-values for overlap in differentially expressed genes were calculated with Fisher's exact test.

---

sistently deregulated KEGG pathways were found at the pre-symptomatic stage. This indicates that major transcriptional changes are associated with symptoms and age but not with the expression of expPABPN1 *per se*.

The transcriptional changes in OPMD were significantly similar to aging-regulated genes ($P = 1.1\times10^{-40}$; **Figure 2C**), and high similarity was also found between OPMD-deregulated and aging-regulated KEGG pathways from two independent studies (**Supplementary Table 2**). These analyses suggest that in both OPMD and muscle aging the major age-associated transcriptional changes occur during the fifth decade. These transcriptional changes are significantly similar. However, muscle weakness is found in OPMD and not in age-matching controls. This suggests

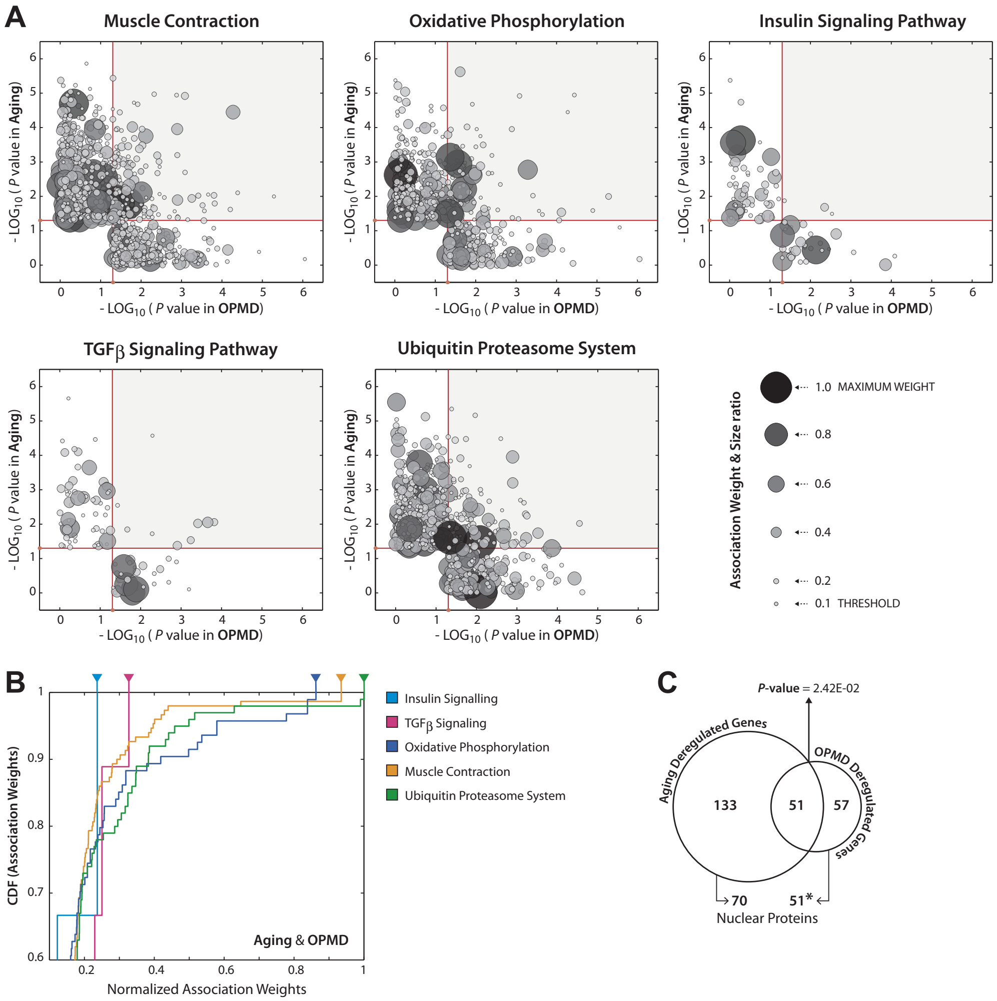


Figure 3 – Analysis of differentially expressed genes in aging and in OPMD reveals that the UPS is the most prominently associated biological process. A) 2D plots of selected biological processes, which are affected in both OPMD (x-axis) and muscle aging (y-axis). Significantly affected genes have *P*-value<0.05 (indicated with red lines). Gene association with 'muscle contraction', 'oxidative phosphorylation', 'insulin signalling pathway', 'TGFβ signalling pathway', and 'ubiquitin-proteasome system' terms is presented by a circle size. Normalized association weights < 0.1 are discarded. **B)** Cumulative distribution function (CDF) plots show the distribution of normalized association weights for overlapping deregulated genes between OPMD and muscle aging (>42 years) for each of the terms in **A**. Arrowheads indicate the maximum association weights.

---

that progression and or amplitude of those transcriptional changes may underlie differences in between OPMD and controls.

The UPS is the most affected pathway in OPMD and muscle aging

Next, we investigated the similarities of molecular changes in OPMD and aging-associated biological pathways using a literature-association study (LAS). In this study, we assessed the association weights of overlapping genes between muscle aging and OPMD with the five most robust aging and OPMD-related pathways: oxidative phosphorylation, insulin signaling, tumor growth factor (TGFβ) signaling, the ubiquitin proteasome system (UPS) and muscle contraction (**Supplementary Table 2**). In the muscle contraction group, the overlapping genes between OPMD and muscle aging had high association weights (**Figure 3A**). This suggests that similar molecular signatures of muscle contraction are found in OPMD and muscle aging. The overlapping genes between OPMD and muscle aging were strongly associated with oxidative phosphorylation and the UPS (**Figure 3A**), while little similarity was found for highly influenced genes in the insulin or TGFβ signaling pathways (**Figure 3A**). This suggests that different key components in insulin or TGFβ signaling pathways are deregulated in OPMD and muscle aging.

The association weights of the overlapping genes in OPMD-deregulated and aging-regulated with five functional groups were ranked in Cumulative Distribution Function (CDF) plots and compared against a theoretical random distribution. The associations of the genes with UPS, oxidative phosphorylation and muscle contraction were much stronger than expected by chance (**Figure 3B**; Kolmogorov-Smirnov test: P = 4.3×10^{-39}, 8.1×10^{-25} and 2.4×10^{-26}, respectively). In contrast, the distribution of association weights for genes in the insulin and TGFβ pathways were insignificant and did not differ from a theoretical random distribution. The low P value is, in part, due to the limited number of overlapping genes between OPMD and aging muscle in the latter pathways. The UPS ranked the highest suggesting that key components of the UPS contribute to both muscle aging and OPMD.

Age- related transcriptional changes are accelerated in OPMD

Clinical muscle weakness in quadriceps is found in OPMD patients but not in age-matching controls (**Supplementary Table 2**). Muscle weakness in quadriceps among healthy subjects is significant in the elderly (Hairi et al., 2010). Therefore, we investigated whether age-dependent expression changes are accelerated in OPMD compared to healthy individuals. Age-dependent expression trends of the probes that differentially expressed in both OPMD and aging were clustered using k-means clustering. One cluster of up- and one cluster of down-regulated probes in aging show earlier and accelerated changes in OPMD carriers (**Figure 4A**). Examples of reprehensive expression trends of individual genes from each cluster are presented in Figure 4B. Among those we identified the cell cycle regulator, *CDKN1A* (p21), and *LMOD1* and *CHRNA1* that are associated with muscle contraction. Among the genes with accelerated expression trends in OPMD, for some the expression is changed at the pre-symptomatic stage. This analysis suggests that expression trends in OPMD change faster compared with controls, and therefore changes in expression profiles are accelerated in OPMD.

Next we evaluated similarities in expression profiles between OPMD and elderly (>80 years). Significant overlap was identified between OPMD-deregulated and elderly-regulated genes ($P = 1.6\times10^{-168}$; **Figure 5A**). From those, 74% showed a similar direction of deregulation. Examples of genes with similar direction of deregulation in both datasets are shown in **Figure 5B**. All genes were identified as aging-regulated in independent studies (Welle et al., 2004; Lu et al., 2004; Rodwell et al., 2004). Since muscle weakness and atrophy is evident in elderly, this analysis suggests that similar molecular changes are associated with muscle weakness in OPMD and elderly.

We also investigated the pool of overlapping genes between OPMD and elderly. The relevance of this gene pool to aging was assessed with the literature concept 'Aging'. The association-weight of these genes to 'Aging' was very strong (**Figure 5C**). This confirms that this procedure can robustly and quantitatively identify gene association to literature concepts. Similar to the pool of overlap-

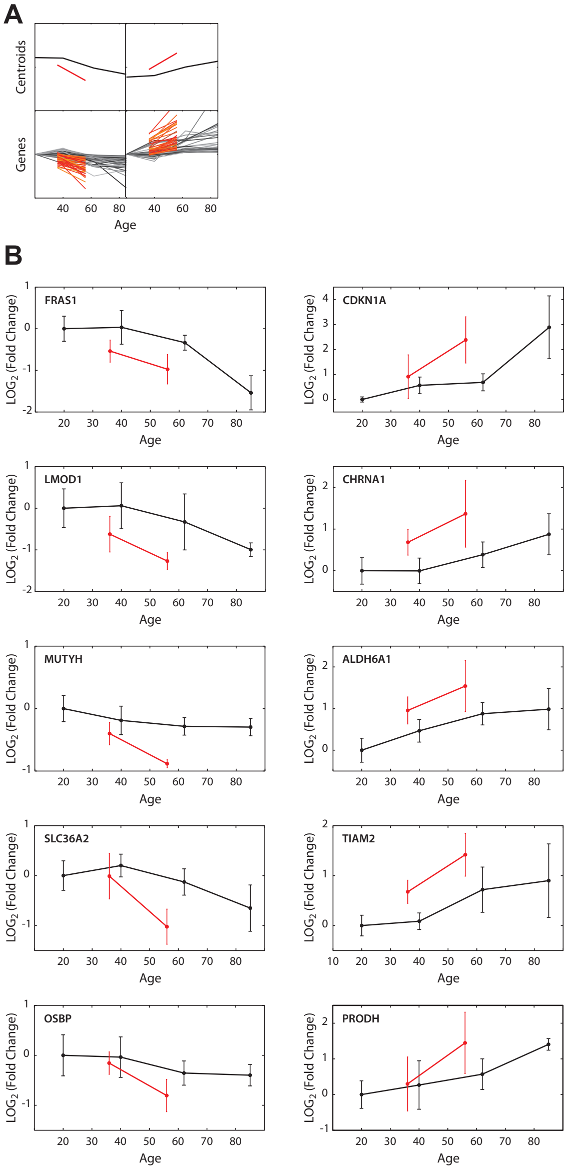


Figure 4 – Aging-associated expression trends are accelerated in OPMD. A) Expression trends of aging (>42)-regulated and OPMD-deregulated probes show progressive transcriptional changes in aging healthy controls (grey lines) and accelerated changes in OPMD (red lines). Upper plots show a summary trend (centroids) of all genes in each cluster, and lower plots show individual genes. **B)** Examples of expression trends of 10 genes from clusters in **A**, in healthy controls (grey lines) and in exPABPN1 carriers at pre-symptomatic and symptomatic stages (red lines). Standard deviations are indicated. Left and right columns show down- or up- regulated expression trends, respectively.

ping genes between aging and OPMD, in the OPMD-elderly pool, strong association was found with oxidative phosphorylation, the UPS and muscle contraction. The association with insulin and TGFβ signalling pathways was less strong (**Figure 5C**).

Protein homeostasis is mainly regulated by the autophagy-lysozyme system and the UPS. Since the UPS ranks the highest in both OPMD-aging and OPMD-elderly pool of genes, we next compared the association-weights of genes associated with lysozyme, autophagy and the UPS in the pool of OPMD-elderly overlapping genes. In contrast to the UPS, the association strength for autophagy and lysosome was very low (**Figure 5C**). The UPS was identified as the most significantly and consistently deregulated pathway in OPMD and models (Anvar et al., 2011). In that study deregulation of genes in autophagy and lysozyme ranked much lower and was not consistently significant in all OPMD model systems. This suggest that deregulation of genes in the UPS has the highest contribution to muscle weakness in both aging and OPMD.

PABPN1 expression progressively declines with aging and the decline is accelerated in OPMD

OPMD is caused by expression of expPABPN1. In the mouse model for OPMD severity of muscle weakness is associated with an increase in aggregates (Davies et al., 2005; Trollet et al., 2010). In models for OPMD aggregation depends on expression level. To our surprise, among the OPMD-deregulated genes in our microarray study we noticed PABPN1. To validate the microarray observation PABPN1 expression levels were determined with RT-qPCR of RNA from *Vastus lateralis*. Expression levels in OPMD patients or expPABPN1 carriers at the pre-symptomatic stage were compared

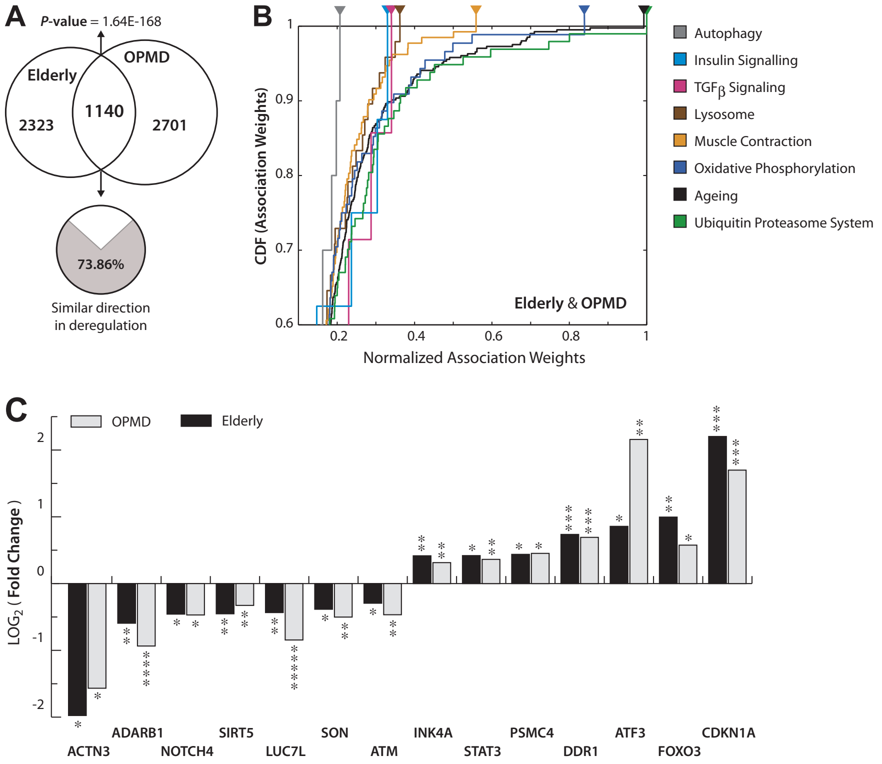


Figure 5 – Similar changes in expression between elderly and OPMD. A) Venn diagram shows the overlap of differentially expressed genes in OPMD- and in elderly (>80 year). From the 1140 overlapping genes, 77% show changes in a similar direction. The P-value for the overlap was calculated with Fisher's exact test. **B)** Histogram show change in expression levels of genes with significantly changed expression in the elderly (80 *vs.* 60 years) and in OPMD patients (*vs.* age-matched controls). All genes are reported in the literature as aging-deregulated (* $P<0.05$, ** $P<0.005$, *** $P<0.0005$, and **** $P<0.00005$). **C)** Cumulative distribution function (CDF) plots show the distribution of normalized association weights for overlapping deregulated genes between OPMD and elderly (>80 years) for each of the terms indicated in the figure. Arrowheads indicate the maximum association weights.

with age-matching control groups. A significant decline in expression was found in OPMD compared with age-matching controls (**Figure 6A**). At the pre-symptomatic stage a slight but insignificant reduction was found (**Figure 6A**). Since OPMD samples are significantly older compared with pre-symptomatic, we next analysed whether a change in PABPN1 expression level is associated with age. RT-qPCR was performed on *Vastus lateralis* from 78 healthy controls aged 17-89. A significant decline in *PABPN1* expression was identified from 43 years onwards (**Figure 6B**). A quadratic model or two linear models describes most accurately the change in PABPN1 expression during age (**Figure 6B**). A significant shift in expression was identified around 43 years (**Table 1**). This age-associated change in *PABPN1* expression shows a similar trend as decline in skeletal muscle strength during aging (Kent-Braun et al., 2002; Roth et al., 2002), which is initiated around midlife and progressively declines onwards. This suggests that changes in PABPN1 expression marks muscle aging. Moreover, symptoms in OPMD, but not the expression of exp-PABPN1 *per se,* are associated with a decline in *PABPN1* expression.

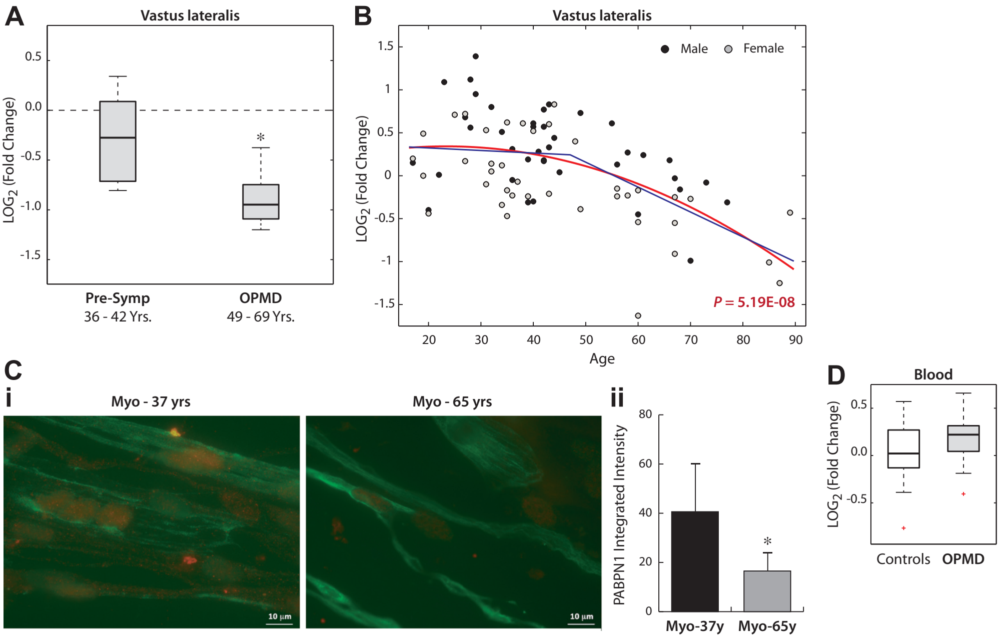


Figure 6 – PABPN1 expression declines in OPMD and during muscle aging. A) Box plot shows *PABPN1* LOG2 fold change in *Vastus lateralis*. Fold change was measured from RT-qPCR and was normalized to *GAPDH* and HRPT genes and to age matching control groups ($N_{\text{pre-symptomatic}}$ = 6, $N_{\text{age-matched control group}}$ = 16; N_{OPMD} = 9. $N_{\text{age-matched control group}}$ = 20). **B)** Scatter plot shows *PABPN1* LOG2 expression in quadriceps of 78 healthy controls between 17 and 89 years. Male and female samples are indicated in black and gray, respectively. A quadratic fit is shown with a red line (age 17-89), gender-corrected *P-value* for the quadratic fit is indicated in red. Blue dashed lines show linear fits for the age groups: 17 - 42 and 43 - 89 years. **C)** PABPN1 protein expression in primary myoblasts from young (37y) and old (65y) donors. **i)** Immunofluorescence of PABPN1 (red) and Desmin (green) in myotube cultures of 37 or 65 year-old donors. Scale bar is 10 mm. **ii)** Histogram shows integrated fluorescence intensity of PABPN1 in myonuclei of 37y and 65y cultures, N_{37y} = 103 and N_{65y} = 87 myonuclei. P value was calculated with the student's T-test, significant difference ($p<0.05$) is indicated with an asterisk. **D)** Box plot shows *PABPN1* LOG2 expression in blood of OPMD patients (N_{OPMD} = 16) and age-matched controls ($N_{\text{age-matched control group}}$ = 12). Expression values were normalized to *GAPDH* and HRPT genes.

To validate the decline in *PABPN1* mRNA expression, PABPN1 protein accumulation was determined in primary muscle cell cultures from 37 or 65 year-old individuals (**Figure 6C**). Protein analysis was performed on cultures that were in vitro propagated for a single passage. A nuclear staining of PABPN1 was found in these myoblasts. A decline in PABPN1 protein accumulation was observed in Myo-65y compared with Myo-37y, whereas the intensity of Desmin staining was unchanged (**Figure 6Ci**). Quantification of nuclear PABPN1 fluorescence intensity in myonuclei of fused myotubes revealed a significant decrease in Myo-65y compared with Myo-37y (**Figure 6Cii**).

PABPN1 is expressed in every cell whilst symptoms in OPMD are predominantly exhibited in a subset of skeletal muscles. To investigate whether the decline in PABPN1 expression is tissue specific, the expression of *PABPN1* was determined in blood samples of OPMD patients. RT-qPCR analysis revealed that *PABPN1* expression levels were unchanged between OPMD patients and age-matching controls (**Figure 6D**). This suggests that a decline in PABPN1 expression in OPMD

Table 1 – Changes in PABPN1 expression depends on chronological age are muscle specific.

Tissue	**Age** (years)	**Beta**	**P-value**
Vastus lateralis	17 – 42 (N = 41)	-0.006 (0.009)	0.37
	43 – 89 (N = 34)	-0.029 (0.006)	<0.0001
Frontal Brain Cortex	26 – 69 (N = 17)	0.002 (0.007)	0.73
	70 – 95 (N = 13)	-0.018 (0.008)	0.04
Blood	42 – 102 (N = 150)	0.001 (0.003)	0.69
Kidney Cortex	27 – 92 (N = 72)	-0.001 (0.002)	0.76
		-0.001 (0.002)	0.42
		-0.003 (0.002)	0.15
Kidney Medulla	29 – 92 (N = 61)	-0.003 (0.002)	0.11
		0.001 (0.002)	0.76
		-0.004 (0.002)	0.06
Rectus Abdominis	24 – 83 (N = 81)	-0.000 (0.003)	0.94
		0.010 (0.007)	0.13
		0.001 (0.003)	0.64
Parotid glands	19 – 71 (N = 13)	0.000 (0.003)	0.93
		0.003 (0.005)	0.64
		-0.001 (0.005)	0.86

Betas (standard errors of the mean) of a linear model are provided per probes. Values for three independent PABPN1 probes are shown for datasets from Kidney cortex, Kidney medulla, Rectus Abdominis and Partotid glands. *P-values* are adjusted for gender. Significant changes are highlighted in bold. N indicates number of samples. Age is indicates in years (y).

is muscle-specific. Next we investigated PABPN1 expression in several aging-related microarray studies from different tissues. A change in PABPN1 expression was not found in Blood, Parotid glands, kidney cortex or kidney medulla (**Table 1**). In postmortal frontal brain cortex we identified a small decline in PABPN1 expression in elderly (**Table 1**). Compared with PABPN1 decline in *Vastus lateralis*, the decline in the brain cortex was smaller and delayed (**Table 1**). Also in *Musculus rectus abdominis* PABPN1 expression was not changed with age (**Table 1**). *Rectus Abdominis* is a typical posture skeletal muscle, while the *Vastus lateralis* is involved in muscle movement. Moreover, muscle weakness is more pronounced in the *Vastus lateralis* compared with *Rectus Abdominis* (Marzani et al., 2005). Together, this analysis suggests that the age-associated decline in PABPN1 expression marks physiological aging in a subset of skeletal muscles.

PABPN1 down-regulation in human muscle cell culture induces cellular senescence and myogenic defects

To investigate the effect of PABPN1 down-regulation in muscle cells, three PABPN1 shRNA clones were selected for functional studies in immortalized human myoblast cultures using the lentivirus expression system. Compared with controls (H1 empty vector and non-transduced cells), the three PABPN1 shRNA clones, 121, 122 and 123, led to a 70%, 40% and 20% decrease in *PABPN1* expression (**Figure 6A**). These clones were selected as they represent a physiological decline in PABPN1. The sh121 clone led to down-regulation that is comparable to the decline in OPMD patients, while the sh122 clone led to a decline as in healthy controls around 60-70 years. The small decline in the sh123-transduced cells was comparable to the expression level in 40-50 year-old controls. Western blot analysis of protein extracts from fused cells confirmed substantial PABPN1 down-regulation in the sh121-transduced cell cultures, and about 40% reduction in

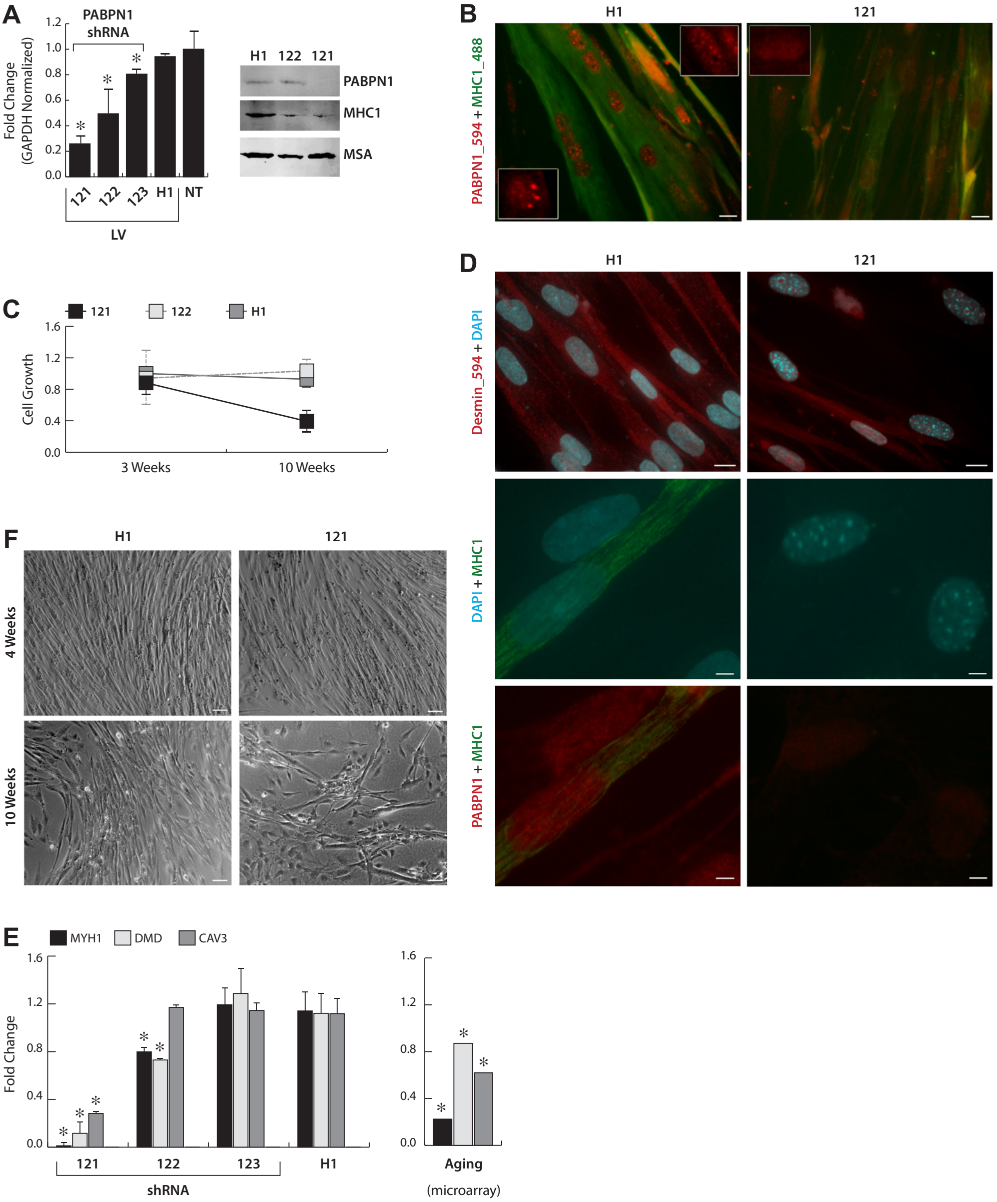


Figure 7 – PABPN1 down-regulation in myotubes shows myogenic defects and cell senescence. Human myotubes were transduced with shRNA specific to PABPN1 (121, 122, and 123) or H1 empty vector. Non-transduced (NT) cells were used as controls. **A)** Histograms show *PABPN1* expression in myoblasts two weeks after transduction. Fold change was normalized to *GAPDH* gene and to non-transduced cells. Averages are of 6 biological replicates. Western blot analysis of PABPN1, MHC1 and MSA in 121-, 122- or H1- transduced myotubes two weeks after transduction. **B)** Immunofluorescence of PABPN1 (labelled with Alexa-594) and myosin (labelled with Alexa-488) in 121- or H1-transduced fused myoblast cultures. Scale bars are 20 mm. A magnification of a single nucleus is shown in the boxed image. **C)** Cell growth analysis of 121-, 122- and H1- transduced myoblasts 3 or 10 weeks in culture. 50,000 cells were plated and were counted after 2 days in culture. Plots show normalized cell number to un-transduced controls. Averages are of 3 biological replicates. **D)** Left: Immunofluorescence of myotube cell cultures of desmin, PABPN1 and

MHC1. Cells were cultured for 10 weeks before fusion. Nuclei were counter stained with DAPI. Scale bars are 15 mm (Desmin) or 5 (PABPN1 and MHC1) mm. **E)** Images of fused myoblast H1- or 121- transduced cultures. Preceding fusion cells were maintained for 4 or 10 weeks after transduction. Scale bar is 30 mm. **F)** Left histogram shows RNA expression of *MYH1*, *DMD*, and *CAV3* in 121-, 122-, 123-, and H1-transduced fused myoblast cultures. Cells were cultured for 3 weeks before fusion. Fold change was normalized to *GAPDH* and to non-transduced cells. Averages are of 3 biological replicates. Significant down-regulation (*P*<0.05) is indicated with asterisks. Right histogram shows Fold change in the microarray study in aging.

sh122-transduced cells (**Figure 7A**). A decrease in the accumulation of nuclear PABPN1 was also verified by immunofluorescence in the sh121-transduced cells. A reduced PABPN1 signal was found in sh121 cells compared with control cells (**Figure 6B**). Nuclear PABPN1 is localized to speckles (Tavanez et al., 2005). In myonuclei of sh121 the speckle localization of PABPN1 was disrupted (**Figure 7B**, box). Together, this demonstrates that shRNAs for PABPN1 induced a decline in mRNA and protein accumulation.

Next we investigated cellular effects of PABPN1 down-regulation. Cell growth was not significantly affected in all myoblast cultures two or three passages after transduction (**Figure 7C**). However, after a longer culturing period, a 60% decline in cell growth was found in the sh121-transduced cells, whereas changes were not found in sh122, sh123-transduced cells or in controls (**Figure 7C**). Senescent cells are marked by heterochromatic foci (HF) (Spector and Gasser, 2003). We observed HF in the sh121-transduced cells but not in controls (**Figure 7D**). PABPN1 expression was undetectable in nuclei with HF (**Figure 7D**). In vivo, the majority of muscle cells are post-mitotic; therefore we compared the abundance of HF nuclei between myoblast and myotube cultures. 24% of myonuclei in 121-fused cultures contained HF whereas in 121-myoblasts only 9% of the cells were with HF. This suggests that the effect of PABPN1 down-regulation on cellular senescence is more pronounced in post-mitotic cells. Senescent muscle cells exhibit reduced fusion (Bigot et al., 2008). The fusion index in control cells was around 70% in transduced cells and controls, and was not significantly affected during in vitro propagation (**Figure 7E**). However, during in vitro propagation of the sh121-transduced cells cell fusion was reduced to 30% (**Figure 7E**). In concordance with cell growth, no significant reduced cell fusion was found in sh122- or sh123- transduced cells. Fusion defects can be associated with reduced expression of sarcomere encoding genes. RT-qPCR of *MYH1*, *DMD* and *CAV3* revealed a significant reduction in fused cultures of sh121-transduced cells (**Figure 7F**). For these genes a significant decline in expression was found in our microarray study (**Figure 7**). The decline in *MHY1* on mRNA level was consistent with a reduced protein accumulation in myotubes (**Figure 7A**). In the sh122- and sh123- transduced cells a gradual decrease in the expression of *MYH1* was observed, which corresponds to the decline in *PABPN1* expression (**Figure 7F**). The expression of *DMD* was significantly affected in the sh122- but not in the sh123- transduced cells. The expression of *CAV3* reduced only in the sh121-transduced cells. Our experiments in this cell model suggest a regulatory role for PABPN1 expression level in induction of cell senescence in muscle cells, which is associated with a gradual change in expression of sarcomeric genes.

DISCUSSION

PABPN1 regulates poly(A) tail length and mRNA stability (Lemay et al., 2010; Kuhn et al., 2009), and thus plays an indispensable role in cell homeostasis by affecting genome-wide mRNA accumulation. Previous studies demonstrated that a complete knockdown of PABPN1 causes shorting of poly(A) tail, which is associated with myogenic defects, including reduction in cell growth and fusion (Apponi et al., 2010; Chartier et al., 2006; Davies et al., 2006; Trollet et al., 2010). Here, for the first time, a significant decline of PABPN1 expression in affected muscles of OPMD patients is

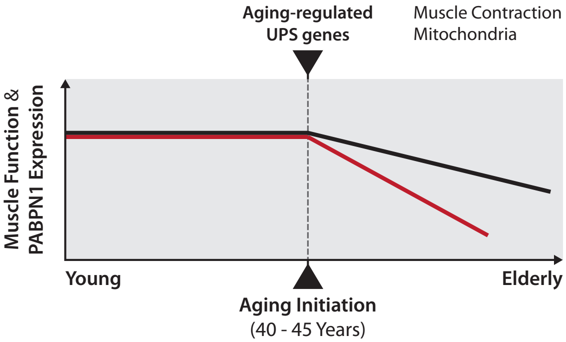


Figure 8 – Schematic presentation of decline in PABPN1 expression in association with protein aggregation during aging. Upper panel represents age-associated changes in PABPN1 expression manifested during midlife, with acceleration in OPMD (in red). Lower panel illustrates the decline in the level of soluble PABPN1 during aging of skeletal muscle, regulated by the UPS.

reported. Since a decline in PABPN1 expression was not found at the pre-symptomatic stage it suggests that the decline in PABPN1 expression is not caused by the expression of expPABPN1, *per se*. We show that a down-regulation of PABPN1 expression, to levels that are found *in vivo* (OPMD and aging), in muscle cell culture leads to cellular defects, including cell senescence and myogenic defects. In accordance with disease progression, the decline in cell growth and fusion correlates with levels of PABPN1 down-regulation. Primary myoblast cultures from OPMD patients also exhibit reduced cell growth and fusion defects (Perie et al., 2006). Overexpression of PABPN1 also leads to muscle cell defects and atrophy, which is associated with genome-wide transcriptional chances (Trollet et al., 2010). Mild overexpression of either expPABPN1 or the wild type allele in fused muscle cell culture also leads to transcriptional changes (Raz et al., 2011b). These changes, however, are significantly smaller compared with high overexpression situations (Raz et al., 2011b). Since PABPN1 regulates poly(A) length and hence mRNA stability, these studies together suggest that manipulations of PABPN1 expression levels below or above a narrow threshold leads to widespread transcriptional changes in muscle cells.

PABPN1 is ubiquitously expressed but symptoms in OPMD are predominately exhibited in a subset of skeletal muscles. Here we found that in OPMD PABPN1 expression declines in skeletal muscles but not in blood. During normal muscle aging, PABPN1 level also decreases. However, this decline is slower and smaller than in OPMD. The decline in PABPN1 expression was not found in other tissues like kidney, Parotid glands, blood or *Rectus Abdominis* muscles, which is less affected during aging. A smaller and delayed decline in PABPN1 was identified in brain cortex. This suggests that a decline in PABPN1 expression is more prominent in skeletal muscles. The decline was progressive from the age 43± years, and perfectly fit to the decline in muscle weakness during aging (Beenakker et al., 2010). Previous studies demonstrated significant muscle weakness in quadriceps of elderly (Kent-Braun et al., 2002; Roth et al., 2002). A major switch in expression profiles in both OPMD and aging was identified during the first half of the fifth decade. This suggests that similar mechanisms initiate muscle weakness in aging and OPMD. Transcriptional similarities between OPMD and elderly suggest differences in progression of aging-regulated muscle weakness between OPMD and normal aging (**Figure 8**).

Protein aggregation is the hallmark of OPMD. Both wild type and mutant PABPN1 are prone to aggregation. However, aggregation potency of expPABPN1 is higher than that of the wild type protein (Raz et al., 2011b). In contrast to the aggregation process of wild type PABPN1, that of expPABPN1 is irreversible and encompasses stable pre-aggregated forms or oligomers (Raz et al., 2011a). Aggregates of both wild type and expPABPN1 entrap a broad rage of nuclear proteins, including components of the UPS (Calado et al., 2000; Anvar et al., 2011). The rate of protein entrapment differs between aggregation process of wild type and mutant PABPN1 (Raz et al., 2011a). Protein entrapment can be associated with transcriptional changes of nuclear proteins and UPS encoding genes (Corbeil-Girard et al., 2005; Anvar et al., 2011). Since proteostasis of nuclear proteins is predominantly regulated by the UPS, changes in expression of UPS encoding genes would affect the ratio of soluble to aggregated proteins. PABPN1 aggregation reduces the levels of soluble PABPN1 (Raz et al., 2011b), and therefore could lead to a similar effect as down-regulation. Aggregation of PABPN1 is regulated by the UPS (Raz et al., 2011b). Moreover, transcriptional changes of the UPS were identified in OPMD and aging. In elderly and OPMD the UPS ranked with a highest association. Functional decline of the UPS is associated with an accumulation and aggregation of misfolded proteins (Balch et al., 2008; Morimoto, 2008; Sherman and Goldberg, 2001). In *C. elegance*, aging is associated with widespread accumulation of aggregated proteins (David et al., 2010). Changes in proteasome activity in skeletal muscles were observed in muscle aging (Ferrington et al., 2005). We suggest that age-associated changes in UPS expression play a role in OPMD onset (**Figure 8**).

Altogether, our data reveals a strong association between PABPN1 expression in OPMD and in muscle aging. A decline in PABPN1 expression marks muscle aging and we suggest that PABPN1 plays an indispensable role in muscle homeostasis. From this study new regulators of aging cells could be identified in future studies.

MATERIALS AND METHODS

Human materials, RNA extraction and RT-qPCR

Datasets: Human and mouse samples that were used in the microarray studies have been previously published (Anvar et al., 2011; Trollet et al., 2010). A summary of human samples is listed in **Supplementary Table 1**.

All human muscle biopsies presented in this study were collected at Radboud Hospital, Nijmegen, Canisius-Wilhelmina Hospital, Nijmegen, The Netherlands, and Rigshospitalet, Denmark, after an approval of the medical ethical committee Arnhem-Nijmegen (CMO nr. 2005/189) and of the local ethical committee, from The NL and Denmark, respectively. OPMD patients and pre-symptomatic were genetically confirmed and underwent clinical investigation including MRC score prior to sampling of muscle biopsy. All quadriceps biopsies were collected using the Bergstrom needle procedure. The biopsies froze immediately in liquid nitrogen and stored at -80 before RNA extraction.

RNA extraction and RT-qPCR were performed as described in (Trollet et al., 2010). Expression levels were calculated according to the ΔΔCT method, and were first normalized to *GAPDH* housekeeping gene and then to controls (17 - 25 years) in the aging studies, or to the age-matching controls in the studies of expPABPN1 carriers. The statistical significance was determined with the Student's t-test. The list of primers used in this study is provided in **Supplementary Table 3**.

Microarray and Statistical Analyses

The human and mouse microarray datasets are publicly available at GEO repository under the accession numbers GSE26605 and GSE26604, respectively. In all datasets genome-wide expression profiles of skeletal muscles from OPMD were compared to controls. *PABPN1* expression in non-muscle tissues was identified from previously published microarrays, all are publically available: frontal cortex: (GEO-GD707, GEO-GSE1572; Lu et al., 2004), *Rectus abdominis* (GEO-GSE5086; Zahn et al., 2006), blood (GEO-GSE16717; Passtoors et al., 2012), kidney (Rodwell et al., 2004) and Parotid glands (GEO-GSE8764; Srivastava et al., 2008).

Data Processing: Quantile normalization was applied on the microarray raw dataset and data quality was assessed by the principal component analysis. Differentially expressed genes between two age-groups were identified by applying hierarchical linear model using limma package in R (Smyth, 2004) at a cut-off of 0.05. Furthermore, a list of aging-deregulated genes was filtered for those that could not be confirmed after integration with additional set of control individuals in an independent dataset. The OPMD-deregulated genes in the OPMD mouse model and patients were identified as previously described (Anvar et al., 2011; Trollet et al., 2010). Probe annotation was carried out using illuminaHumanv3BeadID (human) and illuminaMousev1BeadID (mouse) R packages. Statistical significance of gene overlap was carried out with the Fisher's exact test in R.

The principal component analysis (PCA) was applied on the human dataset to identify outliers and to investigate age-associated variations. PCA analysis was performed in Matlab and in R.

For the literature-aided study (LAS) the association weights between genes and each biological process were mined using Anni 2.1 (Jelier et al., 2008b). The association weights were normalized to the scale between 0 and 1, relative to the maximum association weight. Threshold of 0.1 was applied to remove genes with weak association (based on the level of evidential support in literature). In addition, genes with $P > 0.05$ ($-\log_{10} > 1.3$) in muscle aging and OPMD were excluded.

Cumulative Distribution Function (CDF) plots were used to examine the association distribution for deregulated genes in OPMD and muscle aging. The CDF of $Gene_i$ is defined as the proportion of genes with association weight less than or equal to that of $Gene_i$. The Kolmogorov-Smirnov (KS) test was used to identify distributions that significantly differ from a theoretical distribution, threshold of $P < 10^{-3}$. Statistical tests were performed in Matlab.

The k-means clustering was used to identify similar expression trends. The procedure was made with probes. For the control samples an absolute correlation was applied to cluster probes with reciprocal (up or down) trends. However, in order to optimize the clustering arrangements, average Silhouette (S_{avg}) values are calculated for each cluster in Matlab. Clustering arrangement of partitions with S_{avg} <0.6 were reiterated until the criteria has met. Maximum number of clusters was set to 20 to avoid overly complex clustering arrangement due to the size of the set. The cluster centroids were used to provide summarized age-dependent expression patterns for each cluster.

Statistical analyses of linear and quadratic models were carried out with the SPSS software (IMB) and Matlab, and plots were generated in Matlab.

Pathway Analyses: Genes were mapped to KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) for assessment of significant transcriptional deregulation in aging (>42 years) or in OPMD using global test (Goeman et al., 2004; Jelier et al., 2011). DAVID, a functional annotation clustering tool (Dennis, Jr. et al., 2003; Huang et al., 2009), was used for integration and removing redundancy. The previously published datasets of Welle et al. (Welle et al., 2004) were used for replication and independent confirmation of pathway analysis. Subcellular localization was carried out with Gene Ontology. A recent annotation of genes encoding for aggregation-prone proteins (David et al., 2010) was used to map the human homologues genes using HomoloGene (*http://ncbi.nlm.nih.gov/homologene*) and Inparanoid (*http://inparanoid.sbc.su.se*) online databases. The meta-analysis was carried out on 104 microarray datasets from various organisms as described in Jelier et al. (Jelier et al., 2008a).

Cell culture and Lentivirus transduction

The human 7304 immortalized myoblasts were a kind gift from Francesco Muntoni (University College London, UK) and were prepared by Gillian Butler-Browne and Vincent Mouly (Zhu et al., 2007). The 7304 cells were propagated in a medium containing DMEM+20% Fetal Calf Serum supplemented with an equal volume Skeletal Muscle Cell Media (PromoCell, Heidelberg, Germany) at 37 °C under 5% CO_2. Cell fusion was carried out in a medium containing DMEM+5% Horse Serum. Human skeletal primary myoblasts from a 37-year-old (37y) and a 65-year-old (65y) donor (Tebu-bio, Le Perray en Yvelines, France) are described in (Righolt et al., 2011). Cells were propagated for only one or two passages and subsequently were seeded on collagen-coated glass plates for imaging.

The shRNA in lentivirus expression vectors 121 (TRCN0000000121), 122 (TRCN0000000122) and (TRCN0000000123) 123 were obtained from Sigma-Aldrich. An empty vector, H1, was used as a negative control. Lentivirus particles were produced as described in (Raz et al., 2006). Virus transduction was performed with 2mg/ml polybrene. Cells were cultured with viruses (MOI ~25) overnight, followed by medium refreshing. Transduced cells were maintained in the presence of 5mg/ml puromycin. *PABPN1* down-regulation was determined 3 days, 4 weeks and 8 weeks after transduction using RT-qPCR. Down regulation did not change during culturing. In total, 4 independent transduction experiments were performed. Cell fusion and cell growth experiments were carried out in the absence of puromycin. For cell growth analysis 50,000 cells were seeded in triplicates in a 24 well plate and the number of living cells was counted after two days with TC10^{TM} Automated Cell Counter (BioRad Hercules, CA, USA). Cell growth experiments were carried out 3 and 10 weeks after transduction. Cell fusion was carried out 10 weeks after transduction in triplicates and cell fusion index was determined by dividing the number of nuclei in myotubes to the total number of myotubes.

Immunofluorescence and western blot analyses

The analysis of fused cells was carried out on cells seeded on plastics or on collagen-coated glass plates. Immunofluorescence was carried out as described in (Raz et al., 2006). Images were recorded as described in (Raz et al., 2011b). Primary antibodies used were: anti-Myosin MF20 (Sigma-Aldrich, MO, USA); anti-Desmin (1:500; Cell Signalling Technology, MS, USA) and the anti-PABPN1, 3F5 llama single chain antibody (1:1000; Verheesen et al., 2006), recognised with rabbit-anti-VHH (1:2000). The Alexa 488-, Alexa 430- or Alexa 594- conjugated secondary antibodies against primary antibodies were obtained from Molecular Probes (Invitrogen, CA, USA) and used (1:2000). DAPI (Sigma-Aldrich, MO, USA) was used for DNA counterstaining.

Western blot analysis of total proteins that were extracted from fused cells was carried out as described in (Raz et al., 2011b). Primary antibodies were mouse monoclonal anti-muscle actin (MSA) (1:2000) (Novocastra, Newcastle upon Tyne, UK), 3F5 llama single chain antibody (1:1000) recognised with rabbit-anti-VHH (1:2000) and anti-Myosin MF20 (1:500) (Sigma-Aldrich). Detection of the first antibodies was conducted with the Odyssey Infrared Imaging System (LI-COR Biosciences, NE, USA) and suitable secondary antibodies.

Acknowledgement

This work was funded in part by the European Commission (PolyALA LSHM-CT-2005018675) and Muscular Dystrophy Association (nr: 68016). Association Francaise contre les Myopathies (nr: 15 123) and the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reference List

Abu-Baker,A., Messaed,C., Laganiere,J., Gaspar,C., Brais,B., and Rouleau,G.A. (2003). Involvement of the ubiquitin-proteasome pathway and molecular chaperones in oculopharyngeal muscular dystrophy. Hum. Mol. Genet *12*, 2609-2623.

Anvar,S.Y., 't Hoen,P.A., Venema,A., van der Sluijs,B., van,E.B., Snoeck,M., Vissing,J., Trollet,C., Dickson,G., Chartier,A., Simonelig,M., van Ommen,G.J., van der Maarel,S.M., and Raz,V. (2011). Deregulation of the ubiquitin-proteasome system is the predominant molecular pathology in OPMD animal models and patients. Skelet. Muscle *1*, 15.

Apponi,L.H., Leung,S.W., Williams,K.R., Valentini,S.R., Corbett,A.H., and Pavlath,G.K. (2010). Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. Hum. Mol. Genet. *19*, 1058-1065.

Balch,W.E., Morimoto,R.I., Dillin,A., and Kelly,J.W. (2008). Adapting proteostasis for disease intervention. Science *319*, 916-919.

Beenakker,K.G., Ling,C.H., Meskers,C.G., de Craen,A.J., Stijnen,T., Westendorp,R.G., and Maier,A.B. (2010). Patterns of muscle strength loss with age in the general population and patients with a chronic inflammatory state. Ageing Res. Rev *9*, 431-436.

Benoit,B., Mitou,G., Chartier,A., Temme,C., Zaessinger,S., Wahle,E., Busseau,I., and Simonelig,M. (2005). An essential cytoplasmic function for the nuclear poly(A) binding protein, PABP2, in poly(A) tail length control and early development in Drosophila. Dev. Cell *9*, 511-522.

Berciano,M.T., Villagra,N.T., Ojeda,J.L., Navascues,J., Gomes,A., Lafarga,M., and Carmo-Fonseca,M. (2004). Oculopharyngeal muscular dystrophy-like nuclear inclusions are present in normal magnocellular neurosecretory neurons of the hypothalamus. Hum. Mol. Genet *13*, 829-838.

Bigot,A., Jacquemin,V., Debacq-Chainiaux,F., Butler-Browne,G.S., Toussaint,O., Furling,D., and Mouly,V. (2008). Replicative aging down-regulates the myogenic regulatory factors in human myoblasts. Biol. Cell *100*, 189-199.

Brais,B., Bouchard,J.P., Xie,Y.G., Rochefort,D.L., Chretien,N., Tome,F.M., Lafreniere,R.G., Rommens,J.M., Uyama,E., Nohira,O., Blumen,S., Korczyn,A.D., Heutink,P., Mathieu,J., Duranceau,A., Codere,F., Fardeau,M., and Rouleau,G.A. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet *18*, 164-167.

Brais,B., Xie,Y.G., Sanson,M., Morgan,K., Weissenbach,J., Korczyn,A.D., Blumen,S.C., Fardeau,M., Tome,F.M., Bouchard,J.P., and . (1995). The oculopharyngeal muscular dystrophy locus maps to the region of the cardiac alpha and beta myosin heavy chain genes on chromosome 14q11.2-q13. Hum. Mol. Genet *4*, 429-434.

Calado,A., Tome,F.M., Brais,B., Rouleau,G.A., Kuhn,U., Wahle,E., and Carmo-Fonseca,M. (2000). Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. Hum. Mol. Genet *9*, 2321-2328.

Catoire,H., Pasco,M.Y., Abu-Baker,A., Holbert,S., Tourette,C., Brais,B., Rouleau,G.A., Parker,J.A., and Neri,C. (2008). Sirtuin inhibition protects from the polyalanine muscular dystrophy protein PABPN1. Hum. Mol. Genet. *17*, 2108-2117.

Chartier,A., Benoit,B., and Simonelig,M. (2006). A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. EMBO J *25*, 2253-2262.

Chartier,A., Raz,V., Sterrenburg,E., Verrips,C.T., van der Maarel,S.M., and Simonelig,M. (2009). Prevention of oculopharyngeal muscular dystrophy by muscular expression of Llama single-chain intrabodies in vivo. Hum. Mol. Genet. *18*, 1849-1859.

Corbeil-Girard,L.P., Klein,A.F., Sasseville,A.M., Lavoie,H., Dicaire,M.J., Saint-Denis,A., Page,M., Duranceau,A., Codere,F., Bouchard,J.P., Karpati,G., Rouleau,G.A., Massie,B., Langelier,Y., and Brais,B. (2005). PABPN1 overexpression leads to upregulation of genes encoding nuclear proteins that are sequestered in oculopharyngeal muscular dystrophy nuclear inclusions. Neurobiol. Dis. *18*, 551-567.

David,D.C., Ollikainen,N., Trinidad,J.C., Cary,M.P., Burlingame,A.L., and Kenyon,C. (2010). Widespread protein aggregation as an inherent part of aging in C. elegans. PLoS. Biol. *8*, e1000450.

Davies,J.E., Sarkar,S., and Rubinsztein,D.C. (2006). Trehalose reduces aggregate formation and delays pathology in a transgenic mouse model of oculopharyngeal muscular dystrophy. Hum. Mol. Genet *15*, 23-31.

Davies,J.E., Wang,L., Garcia-Oroz,L., Cook,L.J., Vacher,C., O'Donovan,D.G., and Rubinsztein,D.C. (2005). Doxycycline attenuates and delays toxicity of the oculopharyngeal muscular dystrophy mutation in transgenic mice. Nat Med. *11*, 672-677.

Dennis,G., Jr., Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C., and Lempicki,R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, 3.

Ferrington,D.A., Husom,A.D., and Thompson,L.V. (2005). Altered proteasome structure, function, and oxidation in aged muscle. FASEB J. *19*, 644-646.

Giresi,P.G., Stevenson,E.J., Theilhaber,J., Koncarevic,A., Parkington,J., Fielding,R.A., and Kandarian,S.C. (2005). Identification of a molecular signature of sarcopenia. Physiol Genomics *21*, 253-263.

Goeman,J.J., van de Geer,S.A., de,K.F., and van Houwelingen,H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. *20*, 93-99.

Hairi,N.N., Cumming,R.G., Naganathan,V., Handelsman,D.J., Le Couteur,D.G., Creasey,H., Waite,L.M., Seibel,M.J., and Sambrook,P.N. (2010). Loss of muscle strength, mass (sarcopenia), and quality (specific force) and its relationship with functional limitation and physical disability: the Concord Health and Ageing in Men Project. J. Am. Geriatr. Soc. *58*, 2055-2062.

Huang,d.W., Sherman,B.T., and Lempicki,R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. *4*, 44-57.

Jelier,R., 't Hoen,P.A., Sterrenburg,E., den Dunnen,J.T., van Ommen,G.J., Kors,J.A., and Mons,B. (2008a). Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. BMC. Bioinformatics. *9*, 291.

Jelier,R., Goeman,J.J., Hettne,K.M., Schuemie,M.J., den Dunnen,J.T., and 't Hoen,P.A. (2011). Literature-aided interpretation of gene expression data with the weighted global test. Brief. Bioinform.

Jelier,R., Schuemie,M.J., Veldhoven,A., Dorssers,L.C., Jenster,G., and Kors,J.A. (2008b). Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. *9*, R96.

Kent-Braun,J.A., Ng,A.V., Doyle,J.W., and Towse,T.F. (2002). Human skeletal muscle responses vary with age and gender during fatigue due to incremental isometric exercise. J. Appl. Physiol *93*, 1813-1823.

Kirkwood,T.B. (2005). Understanding the odd science of aging. Cell *120*, 437-447.

Kirkwood,T.B. and Austad,S.N. (2000). Why do we age? Nature *408*, 233-238.

Kuhn,U., Gundel,M., Knoth,A., Kerwitz,Y., Rudel,S., and Wahle,E. (2009). Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J. Biol. Chem. *284*, 22803-22814.

Lemay,J.F., D'Amours,A., Lemieux,C., Lackner,D.H., St-Sauveur,V.G., Bahler,J., and Bachand,F. (2010). The nuclear poly(A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. Mol. Cell *37*, 34-45.

Lexell,J., Taylor,C.C., and Sjostrom,M. (1988). What is the cause of the ageing atrophy? Total number, size and proportion of different fiber types studied in whole vastus lateralis muscle from 15- to 83-year-old men. J. Neurol. Sci. *84*, 275-294.

Lindle,R.S., Metter,E.J., Lynch,N.A., Fleg,J.L., Fozard,J.L., Tobin,J., Roy,T.A., and Hurley,B.F. (1997). Age and gender comparisons of muscle strength in 654 women and men aged 20-93 yr. J. Appl. Physiol *83*, 1581-1587.

Ling,C.H., Taekema,D., de Craen,A.J., Gussekloo,J., Westendorp,R.G., and Maier,A.B. (2010). Handgrip strength and mortality in the oldest old population: the Leiden 85-plus study. CMAJ. *182*, 429-435.

Liu,C.J. and Latham,N. (2011). Can progressive resistance strength training reduce physical disability in older adults? A meta-analysis study. Disabil. Rehabil. *33*, 87-97.

Lu,T., Pan,Y., Kao,S.Y., Li,C., Kohane,I., Chan,J., and Yankner,B.A. (2004). Gene regulation and DNA damage in the ageing human brain. Nature *429*, 883-891.

Marzani,B., Felzani,G., Bellomo,R.G., Vecchiet,J., and Marzatico,F. (2005). Human muscle aging: ROS-mediated alterations in rectus abdominis and vastus lateralis muscles. Exp. Gerontol. *40*, 959-965.

Morimoto,R.I. (2008). Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging. Genes Dev. *22*, 1427-1438.

Passtoors,W.M., Boer,J.M., Goeman,J.J., Akker,E.B., Deelen,J., Zwaan,B.J., Scarborough,A., Breggen,R., Vossen,R.H., Houwing-Duistermaat,J.J., Ommen,G.J., Westendorp,R.G., Heemst,D., Craen,A.J., White,A.J., Gunn,D.A., Beekman,M., and Slagboom,P.E. (2012). Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. PLoS. One. *7*, e27759.

Perie,S., Mamchaoui,K., Mouly,V., Blot,S., Bouazza,B., Thornell,L.E., St Guily,J.L., and Butler-Browne,G. (2006). Premature proliferative arrest of cricopharyngeal myoblasts in oculo-pharyngeal muscular dystrophy: Therapeutic perspectives of autologous myoblast transplantation. Neuromuscul Disord *16*, 770-781.

Raz,V., Abraham,T., van Zwet,E.W., Dirks,R.W., Tanke,H.J., and van der Maarel,S.M. (2011a). Reversible aggregation of PABPN1 pre-inclusion structures. Nucleus. *2*, 208-218.

Raz,V., Carlotti,F., Vermolen,B.J., van der,P.E., Sloos,W.C., Knaan-Shanzer,S., de Vries,A.A., Hoeben,R.C., Young,I.T., Tanke,H.J., Garini,Y., and Dirks,R.W. (2006). Changes in lamina structure are followed by spatial reorganization of heterochromatic regions in caspase-8-activated human mesenchymal stem cells. J Cell Sci. *119*, 4247-4256.

Raz,V., Routledge,S., Venema,A., Buijze,H., van der Wal,E., Anvar,S.Y., Straasheijm,K.R., Klooster,R., Antoniou,M., and van der Maarel,S.M. (2011b). Modeling Oculopharyngeal Muscular Dystrophy in Myotube Cultures Reveals Reduced Accumulation of Soluble Mutant PABPN1 Protein. Am. J. Pathol.

Righolt,C.H., van 't Hoff,M.L., Vermolen,B.J., Young,I.T., and Raz,V. (2011). Robust nuclear lamina-based cell classification of aging and senescent cells. Aging (Albany. NY) *3*, 1192-1201.

Rodwell,G.E., Sonu,R., Zahn,J.M., Lund,J., Wilhelmy,J., Wang,L., Xiao,W., Mindrinos,M., Crane,E., Segal,E., Myers,B.D., Brooks,J.D., Davis,R.W., Higgins,J., Owen,A.B., and Kim,S.K. (2004). A transcriptional profile of aging in the human kidney. PLoS. Biol. *2*, e427.

Roth,S.M., Ferrell,R.E., Peters,D.G., Metter,E.J., Hurley,B.F., and Rogers,M.A. (2002). Influence of age, sex, and strength training on human muscle gene expression determined by microarray. Physiol Genomics *10*, 181-190.

Sahin,E. and Depinho,R.A. (2010). Linking functional decline of telomeres, mitochondria and stem cells during ageing. Nature *464*, 520-528.

Sherman,M.Y. and Goldberg,A.L. (2001). Cellular defenses against unfolded proteins: a cell biologist thinks about neurodegenerative diseases. Neuron *29*, 15-32.

Smyth,G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Spector,D.L. and Gasser,S.M. (2003). A molecular dissection of nuclear function. Conference on the dynamic nucleus: questions and implications. EMBO Rep. *4*, 18-23.

Srivastava,A., Wang,J., Zhou,H., Melvin,J.E., and Wong,D.T. (2008). Age and gender related differences in human parotid gland gene expression. Arch. Oral Biol. *53*, 1058-1070.

Tavanez,J.P., Calado,P., Braga,J., Lafarga,M., and Carmo-Fonseca,M. (2005). In vivo aggregation properties of the nuclear poly(A)-binding protein PABPN1. RNA. *11*, 752-762.

Taylor,E.W. (1915). Progressive vagus-glossopharyngeal paralysis with ptosis: contribution the group of family diseases. J Nerv Ment Dis *42*, 129-139.

Tome,F.M. and Fardeau,M. (1980). Nuclear inclusions in oculopharyngeal dystrophy. Acta Neuropathol. *49*, 85-87.

Trollet,C., Anvar,S.Y., Venema,A., Hargreaves,I.P., Foster,K., Vignaud,A., Ferry,A., Negroni,E., Hourde,C., Baraibar,M.A., ‘t Hoen,P.A., Davies,J.E., Rubinsztein,D.C., Heales,S.J., Mouly,V., van der Maarel,S.M., Butler-Browne,G., Raz,V., and Dickson,G. (2010). Molecular and phenotypic characterization of a mouse model of oculopharyngeal muscular dystrophy reveals severe muscular atrophy restricted to fast glycolytic fibres. Hum. Mol. Genet. *19*, 2191-2207.

van der Sluijs,B.M., van Engelen,B.G., and Hoefsloot,L.H. (2003). Oculopharyngeal muscular dystrophy (OPMD) due to a small duplication in the PABPN1 gene. Hum. Mutat. *21*, 553.

Verheesen,P., de Kluijver,A., van Koningsbruggen,S., de Brij,M., de Haard,H.J., van Ommen,G.J., van der Maarel,S.M., and Verrips,C.T. (2006). Prevention of oculopharyngeal muscular dystrophy-associated aggregation of nuclear polyA-binding protein with a single-domain intracellular antibody. Hum. Mol. Genet *15*, 105-111.

Welle,S., Brooks,A.I., Delehanty,J.M., Needler,N., Bhatt,K., Shah,B., and Thornton,C.A. (2004). Skeletal muscle gene expression profiles in 20-29 year old and 65-71 year old women. Exp. Gerontol. *39*, 369-377.

Welle,S., Brooks,A.I., Delehanty,J.M., Needler,N., and Thornton,C.A. (2003). Gene expression profile of aging in human muscle. Physiol Genomics *14*, 149-159.

Zahn,J.M., Sonu,R., Vogel,H., Crane,E., Mazan-Mamczarz,K., Rabkin,R., Davis,R.W., Becker,K.G., Owen,A.B., and Kim,S.K. (2006). Transcriptional profiling of aging in human muscle reveals a common aging signature. PLoS. Genet. *2*, e115.

Zhu,C.H., Mouly,V., Cooper,R.N., Mamchaoui,K., Bigot,A., Shay,J.W., Di Santo,J.P., Butler-Browne,G.S., and Wright,W.E. (2007). Cellular senescence in human myoblasts is overcome by human telomerase reverse transcriptase and cyclin-dependent kinase 4: consequences in aging muscle and therapeutic strategies for muscular dystrophies. Aging Cell *6*, 515-523.

APPENDIX

Supplementary Table 1 – A list of muscle biopsies of OPMD patients and controls. All expPABPN1 carriers were confirmed by sequence analysis.

Controls						**expPABPN1 carriers**					
						Pre-symptomatic			**Symptomatic**		
Sex	**Age**	**Sex**	**Age**	**Sex**	**Age**	**Sex**	**Age**	**MRC**	**Sex**	**Age**	**MRC**
Female	17	Male	36	Male	55	Female	37	5	Female	49	4
Male	17	Female	36	Female	56	Female	37	5	Female	54	5
Female	19	Male	36	Female	56	Male	38	5	Female	57	4
Female	19	Female	37	Male	56	Female	39	5	Male	59	5
Male	20	Female	38	Male	58	Female	39	5	Female	60	4
Female	20	Male	39	Female	58	Female	41	5	Female	60	4.5
Male	22	Female	39	Female	60				Male	66	4.5
Male	23	Male	39	Female	60				Male	68	3.5
Female	25	Male	40	Female	60				Female	69	4.5
Female	27	Male	40	Male	60						
Male	27	Male	40	Male	61						
Female	27	Female	40	Male	66						
Male	28	Male	41	Female	67						
Male	28	Male	42	Male	67						
Male	29	Male	42	Female	67						
Male	29	Male	42	Female	67						
Female	31	Male	42	Male	68						
Female	31	Male	43	Female	70						
Male	32	Male	43	Male	70						
Female	32	Female	43	Male	73						
Female	32	Female	43	Male	77						
Female	34	Female	44	Female	85						
Female	34	Male	44	Female	87						
Male	34	Male	45	Female	89						
Female	35	Female	48								
Female	35	Male	49								
Female	35	Female	49								

MRC score is a non-linear clinical measure for muscle weakness. MRC in left and right quadriceps was determined at the same day when biopsies were sampled. Values show an average of both sides. MRC in age-matching controls and in pre-symptomatic is 5. 5=normal muscle strength; <5 indicates muscle weakness.

Supplementary Table 2 - Functional Analysis of ageing and OPMD associated transcriptional changes. Biological processes are defined and clustered according to KEGG. Number of genes per pathway is depicted by #. P-value is indicated by P. The proportion of deregulated genes are depicted by %. Significant deregulation of KEGG pathways in OPMD animal models are indicated.

			Muscle Aging						OPMD			
			Human Welle et al. (2004)			Human			Human		M.M.	D.M.
Metabolism			#	*P*	%	#	*P*	%	*P*	%		
	620	Pyruvate metabolism	41	1.1E-07	36.59	40	1.2E-04	37.50	-		*	*
Carbohydrate	51	Fructose and mannose metabolism	35	3.0E-03	20.00	35	2.1E-04	48.57	2.3E-02	28.57	*	*
	562	Inositol phosphate metabolism	49	3.1E-02	16.33	47	8.2E-04	31.91	3.0E-03	17.02	*	*
Energy	190	Oxidative phosphorylation	112	1.9E-05	54.46	108	7.9E-03	28.70	1.4E-02	17.59	*	*
	910	Nitrogen metabolism	24	6.8E-05	25.00	24	2.8E-02	50.00	1.1E-02	20.83	*	*
	564	Glycerophospholipid metabolism	67	3.4E-03	19.40	64	5.0E-07	48.44	1.7E-03	18.75	*	*
Lipid	1030	Glycan structures - biosynthesis 1	118	8.8E-03	23.73	106	4.3E-03	37.74	3.3E-03	14.15	*	*
	590	Arachidonic acid metabolism	54	1.3E-02	14.81	50	2.9E-04	40.00	1.5E-02	18.00	*	
Nucleotide	230	Purine metabolism	145	5.1E-04	23.45	139	5.6E-05	43.88	5.5E-03	28.78	*	*
	240	Pyrimidine metabolism	89	1.6E-03	20.22	86	1.1E-03	36.05	1.9E-02	22.09	*	*
	260	Glycine, serine and threonine metabolism	42	3.6E-06	23.81	39	1.3E-03	30.77	4.4E-02	20.51	*	*
Amino Acid	251	Glutamate metabolism	24	1.1E-04	29.17	23	4.1E-03	47.83	4.2E-02	17.39	*	*
	271	Methionine metabolism	21	1.4E-02	28.57	20	1.2E-02	40.00	4.3E-02	25.00	*	
Glycan Biosynthesis	1032	Glycan structures - degradation	31	2.5E-02	22.58	29	1.7E-02	37.93	3.0E-03	24.14	*	
Cofactors and Vitamins	760	Nicotinate and nicotinamide metabolism	23	1.7E-06	34.78	23	2.2E-04	56.52	6.2E-03	34.78	*	*

Genetic Information Processing

			#	*P*	%	#	*P*	%	*P*	%	M.M.	D.M.
Transcription	3020	RNA polymerase	25	9.3E-04	28.00	22	3.9E-04	36.36	2.1E-02	27.27	*	*
Translation	970	Aminoacyl-tRNA biosynthesis	41	1.2E-03	24.39		-		1.2E-02	28.95	*	*
	3010	Ribosome		-		83	5.9E-03	37.35	-		*	*

			#	*P*	%	#	*P*	%	*P*	%	M.M.	D.M.
Folding, Sorting and Degradation	3050	Proteasome	43	4.7E-04	32.56		-		9.3E-03	34.09	*	*
	4120	Ubiquitin mediated proteolysis	131	1.6E-03	27.48	126	5.4E-06	45.24	1.5E-03	30.16	*	*
	4130	SNARE interactions in vesicular transport	37	1.9E-03	27.03	38	5.8E-03	34.21	3.5E-02	26.32	*	*
Replication and Repair	3030	DNA replication	36	1.4E-03	33.33	34	1.3E-03	41.18	1.5E-02	23.53	*	*
	3420	Nucleotide excision repair	43	2.9E-02	18.60	41	1.6E-03	36.59	8.9E-04	24.39	*	*

Environmental Information Processing

Membrane Transport	2010	ABC transporters - General	41	8.9E-03	24.39	44	2.1E-04	43.18	4.1E-03	25.00	*	*
Signal Transduction	4330	Notch signaling pathway	46	1.1E-05	41.30		-		2.8E-03	18.60	*	*
	4150	mTOR signaling pathway	50	2.4E-05	42.00	49	1.9E-03	44.90	1.7E-02	28.57	*	*
	4370	VEGF signaling pathway	73	1.7E-03	23.29	67	2.5E-06	43.28	5.5E-03	20.90	*	*
	4010	MAPK signaling pathway	265	1.9E-03	24.91	253	7.5E-05	39.13	5.5E-03	18.18	*	N/A
	4012	ErbB signaling pathway	87	2.0E-03	27.59	83	3.1E-06	33.73	2.3E-02	27.71	*	*
	4310	Wnt signaling pathway	148	4.4E-03	25.68	147	1.5E-04	39.46	7.6E-04	21.09	*	*
	4350	TGF-beta signaling pathway	83	7.5E-03	25.30	82	8.2E-05	39.02	9.5E-03	26.83	*	*
	4340	Hedgehog signaling pathway	53	2.0E-02	16.98	56	3.3E-02	26.79	3.7E-03	16.07	*	*
	4070	Phosphatidylinositol signaling system	76	2.2E-02	18.42	75	4.3E-03	34.67	3.4E-03	20.00	*	*
	4630	Jak-STAT signaling pathway	150	2.2E-02	19.33	145	3.9E-05	39.31	5.6E-03	19.31	*	*
	4020	Calcium signaling pathway	176	4.3E-02	21.59	169	3.1E-04	41.42	1.4E-03	18.34	*	*
Signaling Molecules	4512	ECM-receptor interaction	84	3.9E-02	20.24	79	5.3E-03	43.04	5.0E-04	27.85	*	N/A

Cellular Processes

Transport / Catabolism	4140	Regulation of autophagy	34	1.3E-03	26.47	32	2.0E-02	31.25	3.6E-02	18.75	*	*
Cell Motility	4810	Regulation of actin cytoskeleton	210	7.6E-03	22.38	204	9.8E-03	33.33	2.4E-02	16.67	*	N/A

	4110	Cell cycle	111	1.4E-04	28.83	109	5.0E-06	42.20	6.0E-03	21.10	∗	
Cell Growth and Death	4115	p53 signaling pathway	68	1.4E-03	27.94	63	1.1E-06	34.92	-		∗	∗
	4210	Apoptosis	88	4.6E-03	23.86	78	5.4E-06	42.31	1.1E-02	21.79	∗	∗
	4510	Focal adhesion	199	9.0E-04	26.13	190	3.1E-04	44.21	3.4E-03	24.74	∗	N/A
Cell Communication	4520	Adherens junction	74	9.5E-04	28.38	75	1.6E-02	34.67	3.8E-03	25.33	∗	N/A
	4530	Tight junction	129	1.7E-03	22.48	127	1.8E-03	40.16	2.1E-02	21.26	∗	N/A
	4540	Gap junction	94	8.1E-03	23.40	92	1.6E-03	40.22	9.8E-04	31.52	∗	N/A

Organismal Systems

	4660	T cell receptor signaling pathway	94	5.7E-03	23.40	86	7.0E-03	45.35	3.5E-02	15.12	∗	N/A
	4670	Leukocyte transendothelial migration	113	1.3E-02	17.70	107	6.7E-03	45.79	-		∗	N/A
Immune System	4664	Fc epsilon RI signaling pathway	76	1.9E-02	17.11	73	4.1E-04	45.21	1.4E-03	17.81	∗	N/A
	4650	Natural killer cell mediated cytotoxicity	130	2.4E-02	17.69	121	1.5E-03	42.15	3.2E-02	11.57	∗	N/A
	4620	Toll-like receptor signaling pathway	101	3.7E-02	17.82	93	1.2E-02	45.16	1.6E-02	19.35	∗	N/A
	4910	Insulin signaling pathway	136	8.4E-04	30.88	135	3.3E-05	39.26	1.6E-03	29.63	∗	N/A
Endocrine System	4912	GnRH signaling pathway	98	1.1E-02	23.47	95	5.0E-05	40.00	7.4E-04	30.53	∗	N/A
	4916	Melanogenesis	99	1.7E-02	20.20	101	5.5E-03	30.69	6.9E-04	25.74	∗	N/A
Nervous System	4720	Long-term potentiation	70	9.7E-03	25.71	68	5.9E-03	36.76	2.6E-03	25.00	∗	N/A

Human Diseases

Immune System	5340	Primary immunodeficiency	34	1.2E-02	14.71		-		3.5E-02	12.50	∗	N/A
	5010	Alzheimer's disease	158	5.0E-06	46.84	151	6.5E-05	29.80	2.6E-03	21.85	∗	N/A
Neurodegenerative	5012	Parkinson's disease	113	1.4E-05	58.41	104	2.3E-03	27.88	2.1E-02	18.27	∗	N/A
	5040	Huntington's disease	31	2.9E-03	32.26	28	9.6E-03	42.86	5.0E-03	35.71	∗	N/A
	5014	Amyotrophic lateral sclerosis (ALS)	55	6.4E-03	27.27	51	7.5E-05	37.25	1.0E-02	23.53	∗	N/A

			#	*P*	%	#	*P*	%	*P*	%	M.M.	D.M.
Metabolic	4930	Type II diabetes mellitus	42	2.5E-03	35.71	41	7.4E-05	36.59	2.6E-02	19.51	∗	N/A
Infectious	5130	Pathogenic Escherichia coli infection	49	1.3E-03	38.78	48	4.8E-02	45.83	2.3E-02	22.92	∗	N/A
	5110	Vibrio cholerae infection	56	2.4E-03	25.00	57	5.1E-03	45.61	3.7E-03	28.07	∗	N/A
	5120	Epithelial cell signaling in Helicobacter …	65	1.2E-02	20.00	62	1.6E-05	53.23	2.5E-02	29.03	∗	N/A
Cancer	5211	Renal cell carcinoma	69	3.7E-05	34.78	67	1.9E-05	40.30	3.1E-02	20.90	∗	N/A
	5216	Thyroid cancer	29	1.7E-04	41.38	28	4.7E-04	42.86	1.2E-02	25.00	∗	N/A
	5215	Prostate cancer	90	2.3E-04	26.67	84	5.9E-03	45.24	4.7E-03	29.76	∗	N/A
	5218	Melanoma	71	5.3E-04	30.99		-		3.9E-02	19.12	∗	N/A
	5214	Glioma	64	7.9E-04	32.81	62	1.2E-02	38.71	5.4E-03	24.19	∗	N/A
	5220	Chronic myeloid leukemia	74	8.5E-04	36.49	70	3.4E-05	32.86	2.8E-02	22.86	∗	N/A
	5223	Non-small cell lung cancer	54	1.2E-03	29.63	51	1.4E-02	33.33	3.4E-02	17.65	∗	N/A
	5210	Colorectal cancer	83	1.4E-03	31.33	81	2.7E-03	38.27	1.7E-02	20.99	∗	N/A
	5222	Small cell lung cancer	86	2.2E-03	27.91	78	2.7E-04	53.85	1.3E-02	25.64	∗	N/A
	5213	Endometrial cancer	52	2.4E-03	26.92	50	2.3E-03	40.00	3.8E-02	24.00	∗	N/A
	5221	Acute myeloid leukemia	57	4.5E-03	29.82	53	1.8E-05	47.17	3.0E-02	28.30	∗	N/A
	5217	Basal cell carcinoma	53	2.7E-02	18.87	53	4.3E-03	33.96	2.3E-04	16.98	∗	N/A

Supplementary Table 3 – A primer list for RT-qPCR.

Gene	FW Primer	RV Primer
GUSB	5’ CTCATTTGGAATTTTGCCGATT	5’ CCGAGTGAAGATCCCCTTTTTA
GapDH	5’ CAACGAATTTGGCTACAGCA	5’ AGGGGTCTACATGGCAACTG
PABPN1	5’ ATGCCCGTTCCATCTATGTTG	5’ GCCTGGTCTGTTGGTTCGTT
MYH1	5’ TGGACAAACTGCAAGCAAAG	5’ GACCTGGGACTCAGCAATGT
CAV3	5’ CTGTTGCCTGAGCACAAAAA	5’ GTTAGCCAAAGGGGAGGTTC
DMD	5’ TGAGAGCTTTATTGCTGCATTTT	5’ CATGCCATGTGATGTTTATGC

NETWORKS IN BIOLOGY

beyond differential expression

PART TWO

CHAPTER FOUR

The identification of informative genes from multiple datasets with increasing complexity

Seyed Yahya Anvar[1,2,*], Peter A.C. ‘t Hoen[2] and Allan Tucker[1]

In microarray data analysis, factors such as data quality, biological variation, and the increasingly multi-layered nature of more complex biological systems complicates the modelling of regulatory networks that can represent and capture the interactions among genes. We believe that the use of multiple datasets derived from related biological systems leads to more robust models. Therefore, we developed a novel framework for modelling regulatory networks that involves training and evaluation on independent datasets. Our approach includes the following steps: (1) ordering the datasets based on their level of noise and informativeness; (2) selection of a Bayesian classifier with an appropriate level of complexity by evaluation of predictive performance on independent data sets; (3) comparing the different gene selections and the influence of increasing the model complexity; (4) functional analysis of the informative genes. In this paper, we identify the most appropriate model complexity using cross-validation and independent test set validation for predicting gene expression in three published datasets related to myogenesis and muscle differentiation. Furthermore, we demonstrate that models trained on simpler datasets can be used to identify interactions among genes and select the most informative. We also show that these models can explain the myogenesis-related genes (genes of interest) significantly better than others ($P < 0.004$) since the improvement in their rankings is much more pronounced. Finally, after further evaluating our results on synthetic datasets, we show that our approach outperforms a concordance method by Lai *et al.* in identifying informative genes from multiple datasets with increasing complexity whilst additionally modelling the interaction between genes. We show that Bayesian networks derived from simpler controlled systems have better performance than those trained on datasets from more complex biological systems. Further, we present that highly predictive and consistent genes, from the pool of differentially expressed genes, across independent datasets are more likely to be fundamentally involved in the biological process under study. We conclude that networks trained on simpler controlled systems, such as *in vitro* experiments, can be used to model and capture interactions among genes in more complex datasets, such as *in vivo* experiments, where these interactions would otherwise be concealed by a multitude of other ongoing events.

---

1 Center for Intelligent Data Analysis, School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK **2** Center for Human and Clinical Genetics, Leiden University Medical Center, the Netherlands.

* To whom correspondence should be addressed at: s.y.anvar@lumc.nl

BMC Bioinformatics, 2010,
Jan 15; 11:32; doi:10.1186/1471-2105-11-32

BACKGROUND

High-throughput gene expression profiling experiments have increased our understanding of the regulation of biological processes at the transcriptional level. In bacteria (Bockhorst et al., 2003) and lower eukaryotes, such as yeast (Segal et al., 2003), modeling of regulatory interactions between large numbers of proteins in the form of regulatory networks has been successful. A regulatory network represents relationships between genes and describes how the expression level, or activity, of genes can affect the expression of other genes. The network includes causal relationships where the protein product of a gene (e.g. transcription factor) directly regulates the expression of a gene but also more indirect relationships. Modeling has been less successful for more complex biological systems such as mammalian tissues, where models of regulatory networks usually contain many spurious correlations. This is partly attributable to the increasingly multi-layered nature of transcriptional control in higher eukaryotes, e.g. involving epigenetic mechanisms and non-coding RNAs. However, a potential major reason for the decreased performance is due to *biological complexity* of datasets which can be defined as the increase of biological variation and the presence of different cell types, which is not compensated by an increase in the number of replicate data points available for modeling. There is an urgent need to identify regulatory mechanisms with more confidence to avoid wasting laborious and expensive wet-lab follow-up experiments on false positive predictions.

The main paradigms of this paper are that regulatory interactions that are consistently found across multiple datasets are more likely to be fundamentally involved and that these regulatory interactions are easier to find in datasets with less biological variation. In the end, regulatory networks trained on less complex biological systems could thus be used for the modeling of the more complex biological systems. We do this using a novel computational technique that combines Bayesian network learning with independent test set validation (using error and variance measures) and a ranking statistic. Whilst Bayesian networks and Bayesian classifiers have been used with great success in bioinformatics (Friedman et al., 2000; Xu et al., 2004), an important weakness has been that, when trying to build models that reveal genuine underlying biological processes, a highly accurate *predictive* model is not always enough (Grossman and Domingos, 2004). The ability to *generalize* to other datasets is of greater importance (Peña et al., 2005). Simple cross-validation approaches on a single dataset will not necessarily result in a model that reflects the underlying biology and therefore will not generalize well. Our approach is to exploit multiple datasets of increasingly complex systems in order to identify more informative genes reflecting the underlying biology.

Bayesian networks have been an important concept for modeling uncertain systems (Pearl, 1986; Buntine, 1996; Heckerman, 1998; Friedman and Koller, 2003). In the last decade several researchers have examined methods for modeling gene expression datasets based on Bayesian network methodology (Segal et al., 2003; Friedman et al., 2000; Xu et al., 2004). These networks are directed acyclic graphs (DAG) that represent the joint probability distribution of variables efficiently and effectively (Friedman et al., 1997). Each node in the graph represents a gene, and the edges represent conditional independencies between genes. Bayesian networks are popular tools for modeling gene expression data as their structure and parameters can easily be interpreted by biologists.

Bayesian classifiers are a family of Bayesian networks that are specifically aimed to classify cases within a data set through the use of a class node. The simplest is known as the naïve Bayes classifier (NBC) where the distribution for every variable is conditioned upon the class and assumes independence between the variables. Despite this oversimplification, NBCs have been shown to perform very competitively on gene expression data in classification and feature selection problems (Grossman and Domingos, 2004; Fielding, 2007; Tobler et al., 2002). Other Bayesian classifiers, which often have higher *model complexity* as they contain more parameters, involve learning different networks such as trees between the variables and therefore relax the independence assumption (Friedman et al., 1997). The logical conclusion is the general Bayesian Network Classifier (BNC) which simply learns a structure over the variables including the class node. In this paper, we explore the use of the NBC, and the BNC for predicting expression on independent datasets in order to identify informative genes using classifiers of differing complexity.

Accordingly, in order to optimize the classifier and choose the best method, we need to consider the classifiers' bias and variance. Since bias and variance have an inverse relationship (Fielding, 2007), which means decreasing in one increases the other, cross-validation methods can be adopted in order to minimize such an effect. The k-fold cross-validation (Fielding, 2007; Stone, 1974) randomly splits data into k folds of the same size. A process is repeated k times where k-1 folds are used for training and the remaining fold is used for testing the classifier. This process leads to a better classification with lower bias and variance (Kohavi, 1995) than other training and testing methods when using a single dataset. In this paper, we exploit bias and variance using both cross-validation on a single dataset and also independent test data in order to learn models that better represent the true underlying biology. In the next section we provide a description of the gene identification algorithm for identifying gene subsets that are specific to a single simple dataset as well as subsets that exist across datasets of all biological complexity. We used van den Bulcke *et al.* (2006) proposed model for generating synthetic datasets to validate our findings on real microarray data. Moreover, we evaluate the performance of our algorithm by comparing the ability of this model in identifying the informative genes and underlying interactions among genes with the concordance model. Finally, we present the conclusion and summary of our findings in the last section.

METHODS

Multi-Data Gene Identification Algorithm

The algorithm involves taking multiple datasets of increasing biological complexity as input and a repeated training and testing regime. Firstly, this involves a k-fold cross-validation approach on the single simple dataset (from now on we refer to this as the cross-validation data) where Bayesian networks are learnt from the training set and tested on the test set for all k folds. These folding arrangements have been used again for assessing a final model. The Bayesian Network learning algorithm is outlined in the next section.

The Sum Squared Error (SSE) and variance is calculated for all genes over these folds by predicting the measured expression levels of a gene given the measurements taken from others. Next, the same models from each k fold are tested on the other (more complex) datasets (the independent test data) and SSE and variance are again calculated. These SSE and variances are used to rank the genes according to their informativeness (which represents the most predictive and influential genes). Those that are ranked highly in the single-dataset cross-validation experiments will be informative, specific to the single datasets experiment, whereas those that are ranked highly on the independent datasets should be informative in a more general sense in that they are predictive (low SSE) and consistent (low variance) across datasets of all complexity. We evaluate the statistical significance of these rankings using a method proposed by Zhang *et al.* (2006). The full details are outlined in Algorithm 1 where *TrainD* represents the training data (cross-validation

Algorithm 1 - Multi Data Gene Identification Algorithm.

```
Input:  {TrainD, $TestD_1$, …$TestD_M$, folds}
        for $k$ = 1:folds
                Learn BN using Algorithm 2 on training folds of TrainD
                Score SSE on test fold $k$ of TrainD
                Score SSE on all independent test datasets {$TestD_1$…$TestD_M$}
        end for
        Calculate variance of SSE over all $k$ folds on TrainD and {$TestD_1$…$TestD_M$}
        Create gene rankings: trainR_SSE, train_var,
        {$testR\_SSE_1$…$testR\_SSE_M$} and
        {$testR\_var_1$…$testR\_var_M$} by ordering the genes
        on the respective SSE and variance scores

Output: trainR_SSE, train_var,
        {$testR\_SSE_1$…$testR\_SSE_M$}
        {$testR\_var_1$…$testR\_var_M$}
```

data, here the relatively simple datasets), and $TestD_1$ … $TestD_M$ represent the more complex test datasets, independent test data.

Bayesian Network Structure Learning

The goal of learning gene regulatory networks using Bayesian network approaches is to establish the structure of the network and then to parameterize the conditional probability tables (Su and Zhang, 2006). As the number of possible network structures is huge, learning the structure of a network has a high computational cost. Since the effective learning of network structure engages a trade-off of bias vs. variance, the necessity of designing an algorithm in which it can generate an ideal structure for a given dataset, with a degree of biological complexity, is crucial (Chickering et al., 2004). In this study, instead of using well studied but unrealistic and sometimes not effective classifiers such as NBC and Tree Augmented Networks (TAN), we use an optimization approach that uses a simulated annealing search and the Bayes Information Criterion (BIC) as a scoring metric (Schwarz, 1978). The advantage of simulated annealing over other methods (like greedy searches or hill climbing) is that it aims to avoid local maxima (Friedman et al., 1997). We have chosen the BIC as a fitness function as it is less prone to overfitting through the use of a penalizing term for overly complex models.

Bayesian networks with more connections between their nodes require a higher number of parameters and as a result increase the complexity of the models exponentially (Lam and Bacchus, 1994). Therefore, we explore three different classes of model learning: the Selective Naïve Bayes (SNB) where only links between a class node representing differentiation status and a gene are explored, a search that explores structures with links between genes but limiting each gene to having only one parent (1PB). Limiting the number of parents in a Bayesian network is common practise but can be considered a crude approach to reducing parameters. As a result we also explore a full unlimited structure learning (NPB) and learn these structures using the simulated annealing with the BIC scoring metric (which naturally penalises overly complex networks). In this study, the initial state of the structure is an empty DAG with no link. In order to alter the network structures, three operators have been used within the simulated annealing. These operators are adding, removing, or swapping links to generate a new network for validation. These alterations can be either accepted or rejected. The outline of this procedure can be found in Algorithm 2.

Algorithm 2 - Simulated Annealing Structure Learning.

```
Input:  $t_0$, maxfc, D
        $fc=0, t=t_0, t_n=0.001$
        $c=(t_n/t_0)^{1/maxfc}$
        Initial bn to a Bayesian classifier with no inter-gene links
        results = bn
        oldscore=score(bn)
        while fc<maxfc do
            for each operator do
                apply operator to bn
                newscore=score(bn)
                fc=fc+1
                dscore=newscore-oldscore
                if newscore>oldscore then
                    result=nbc
                else if $r(0,1)<e^{dscore/t}$ then
                    Undo the operator
                end if
            end for
            $t=t \times c$
        end while
Output: result
```

Prediction and Ranking

Zhang *et al.* (2006) proposed a method to convert a set of gene rankings into position p-values to evaluate the significance of a given gene. However, this involved working with resampling techniques upon a single dataset. Here, we use the ranking lists according to the model's average SSE and variance for both the original simple dataset and the independent test sets in order to generate position p-values. This requires us to include, a number of random genes which can be counted as uninformative genes. By comparing the actual ranking of the gene with the null distribution we can calculate the position p-values. In this paper we are using three independent datasets so we do not need to use resampling in order to generate more gene rankings as Zhang *et al.* (2006) did in their experiments. In addition, the different rankings will have different interpretations as some are based purely on the simple dataset whilst others are influenced by error and variance on the more biologically complex independent data.

Datasets

With the aim of investigating the influence of the complexity of a gene expression dataset on the performance of classifiers in identifying the gene regulatory network, three gene expression datasets (with increasing biological variation) have been chosen for this study [GSE3858 (Cao et al., 2006), GSE1984 (Iezzi et al., 2004), and GSE989 (Tomczak et al., 2004)]. These three datasets are all concerned with the differentiation of cells into the muscle (Myogenic) lineage. During this process, mononucleated precursor cells stop to proliferate, differentiate and fuse with each other to become elongated multinucleated myotubes or myofibres. This in-vitro system mimics the formation of new muscle fibres in-vivo. The cell types differ between the different datasets:

- GSE3858: Embryonic fibroblasts (EF)
- GSE989 and GSE1984: C2C12 tumor cell line that has the potential for differentiation into different mesodermic lineages (mainly muscle and bone)

Also methods to drive cells into myogenic differentiation differ:

- GSE3858: Exogenous expression of the myogenic transcription factors are Myod and Myog.
- GSE989 and GSE1984: Serum Starvation

Table 1 - Specification of three muscle differentiation datasets.

Dataset	Cell Type	Platform	Samples	Time Points
Tomczak	C2C12	Affy U74A	24	8
Cao	EF	Affy 430.2	36	4
Sartorelli	C2C12	Affy U74A	32	6

In addition, the study by Sartorelli included different treatments that affect the timing and efficiency of the myogenic differentiation process. The time points for sampling differ between the studies (**Table 1**). The class node reflecting the differentiation status had two possible states: undifferentiated (for all time points until myogenic differentiation was induced) and differentiated (for time points where myogenic differentiation had been induced). In the rest of this paper we call these datasets by the name of the first author (e.g. Cao instead of GSE3858).

Data Processing and Analysis

The raw microarray data were normalized and summarized with the RMA method (Irizarry et al., 2003), using the affy package in R. Only the 8904 probesets common to the Affymetrix U74A and 430.2 used in mentioned studies were considered in the analysis. All datasets were standardized to mean 0 and the standard deviation 1 across the genes. For the scope of this paper, first, we selected for each dataset a subset of 100 genes most affected by the induction of differentiation. These genes were identified with Student's t-test which compared samples from undifferentiated and differentiated cell cultures, disregarding the time of differentiation. An additional 50 genes were randomly selected to be able to calculate ranking p-scores described above and using the Kolmogorov-Smirnov test. For cross-validation we divided Cao dataset into 9 folds, Sartorelli into 8 folds, and Tomczak into 6 folds based upon the number of samples in each dataset. Simulated annealing has three attributes which should be set before starting the learning phase. It is crucial to set an appropriate initial temperature, sufficient number of iterations, and a convenient fitness function. In this study, the initial temperature has been set to 10 and it terminates at 0.001. The number of iterations has been set to 1000 for the first set of experiments only using most informative genes (top 100) and then we set the number of iterations to 1500 since we added 50 uninformative genes to the network. The code is implemented in Matlab 2007a using the Bayes Net toolbox (Murphy, 2001) to generate gene regulatory networks.

Analysis of myogenesis-Related genes

Myogenesis-related genes are defined as genes associated with the Gene Ontology term "Muscle Development" supplemented with all genes strongly associated with Myogenesis in the biomedical literature, as determined with the literature analysis tool Anni v2.0 (Jelier et al., 2008) with the association score greater than 0.02.

Analysis of Synthetic datasets

The use of datasets in which the underlying network is known enables us to validate the new algorithms that have been developed to identify gene regulatory networks and capture the most informative genes. van den Bulcke *et al.* (2006) proposed a new methodology to generate synthetic datasets where the network structure is known and biological, experimental, and model complexity can be manipulated. However, a disadvantage of this approach is that the generated networks can contain some overlapping pieces of the known network which may weaken the models being probabilistically independent (Haynes and Brent, 2009). Whilst SynTReN uses resampling from potentially overlapping networks, the generated data undergoes a robust statistical cross-validation regime ensuring that any prediction is applied to unseen data. The focus of this paper is upon the prediction of increasingly complex datasets, sampled from some underlying biological process. Consequently, these synthetic datasets can be used for validating the performance of our methodology in identifying the informative genes and the interactions among them in real microarray data. SynTReN (Van den Bulcke et al., 2006) generates networks with more realistic topological characteristics and since we use this application to investigate the impacts of biological, experimental, and model complexity on identifying informative genes using the same sub-network is an advantage. Three datasets have been generated on the well-described network structure of *E. coli* (Ma et al., 2004) which contains 1330 number of nodes and 2724 interactions. These datasets have been generated in a manner that they can match the key characteristics of real microarray datasets we used in this study (for instance, limiting the number of genes that were selected for modelling to 150). This enables us to investigate the possibility of reproducing similar results on synthetic data which can be easily corrected for differences such as number of samples and time points per dataset (see **Additional file 1**) and avoid weakening the probabilistically independent assumption of the generated datasets.

Analysis of Concordance between datasets

The study of the concordance between microarray datasets has increased considerably in the past few years (Miron et al., 2006). However, a robust statistical method for examining the concordance or discordance among microarray experiments carried out in different laboratories is yet to develop. Methods such as multiplication of gene p-values in order to generate a list of rankings for concordance genes showed bias towards datasets with higher significance level (Rhodes et al., 2002). Lai *et al.* (2009) proposed a promising methodology (which we call concordance model) to investigate the concordance or discordance between two large-scale datasets with two responses. This method uses a list of z-scores, generated using a statistical test of differential expression, as an input to evaluate the concordance or discordance of two datasets by calculating the mixture model based likelihoods and testing the partial discordance against concordance or discordance. Additionally, the statistical significance of a test is being evaluated by the parametric bootstrap procedure and a list of gene rankings is being generated which can be used for integrating two datasets efficiently. In this paper we are using a set of gene rankings generated by this method to evaluate the performance of our model in identifying informative genes from multiple datasets with increasing complexity.

RESULTS

The aim of this study is to demonstrate firstly, the influence of model complexity in discovering accurate gene regulatory networks on multiple datasets with increasing biological complexity. Secondly, to investigate if cleaner and more informative datasets can be used for modelling more complex ones. Therefore, three public datasets that are concerned with the differentiation of cells into muscle lineage were chosen for this study. From a biological point of view, Sartorelli is the most complex dataset since it involves different treatments influencing myogenesis. Tomczak and Cao are less complex datasets. It is difficult to say how their complexity relates since Tomczak uses more heterogeneous stimuli to induce differentiation but has more time points, while Cao uses more defined stimuli (Myod or Myog transduction) and less time points. In order to meet the scope of this study, we evaluated the quality and informativeness of these datasets based on two criteria. Firstly, we calculated the average correlations between replicates as a measurement of noisiness of each dataset. Secondly, using Student's t-test method, we counted the number of differentially expressed genes with the significance levels of 0.05 and 0.01 as a measurement of informativeness (**Table 2**). Although the average correlations between replicates in all three datasets are very close, datasets differ in number of significant genes they hold. Tomczak is the most informative dataset as it includes the most number of significant genes and has a higher average correlation value for the replicate samples in the dataset which represent the lowest level of noise. In contrast, Sartorelli contains the least differentially expressed genes with almost 12% of what Tomczak contains. Moreover, it has the lowest average correlation value and can be marked as the most complex dataset to model in this study as it has the highest noise level and the least number of informative genes. Therefore, we ordered these datasets by increasing biological complexity in the following way: Tomczak, Cao, and Sartorelli.

Table 2 - The average correlations between replicates and number of differentially expressed genes (based on BH corrected p-values) in each dataset.

		Genes with a P-value (BH) less than	
Dataset	**Correlation**	**0.05**	**0.01**
Tomczak	0.975	4602	3604
Cao	0.971	3668	2623
Sartorelli	0.964	1199	458

Comparison of classifiers and network analysis

We now explore how the different classifiers performed on these three datasets. Figure 1 shows the average error rate of the different classifiers trained on each given dataset. It can be seen that of the three classifiers, 1PB and NPB generated the same pattern and have very close error rates on cross-validation (training) sets. However, it is evident that NPB (particularly on Tomczak) performs poorer than 1PB on the independent test set, possibly due to overfitting as these models contain more parameters. Even though SNB performed poorly on both the cross-validation test and the independent data test, in some cases it could compete with NPB which appears to be too complex to predict some of the independent datasets accurately. Hence, 1PB has performed favorably, both in terms of average error rate and the difference between the cross-validation test and the independent data test (see **Additional file 1** for complete set of results).

According to Mac Nally (2000) simple models should be sought for various reasons. Firstly, simple models are more stable and capable of not overfitting to noise in the data which will influence the performance of classifier with future data. Secondly, they tend to provide a better insight into causality and interactions among genes. Finally, reducing the number of parameters will decrease the cost of validating a model for current and future data. However, we need a model that matches the complexity of data sets. Considering this argument along with our first set of results, we chose 1PB as a model that can capture the interactions among genes and does not overfit to noise. In order to understand the impacts of using different datasets for gene selection and training 1PB classifier (which will be discussed in the next section), we need to analyse the performance of the 1PB classifier on the top 100 (most informative) genes in more detail.

Additional file 1, Figure S7 represents the comparison of the error rate of the 1PB classifier on cross-validation versus the independent test. It is shown that the 1PB classifier trained on Tomc-

Figure 1 - The comparison of classifiers with increasing model complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene and an overall SSE on cross-validation set and independent test set are illustrated in this figure.

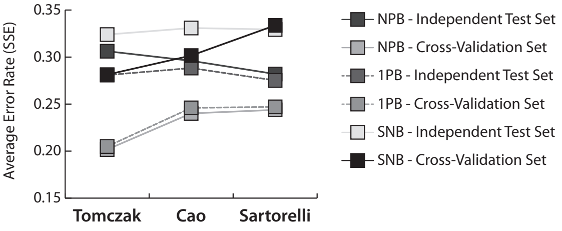
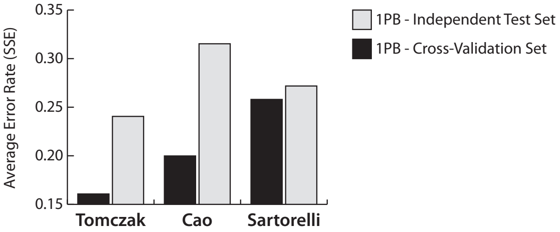


Figure 2 - Evaluating the accuracy of 1PB using different datasets for gene selection. We selected genes using only one dataset (black) at a time and compared the average error rate of 1PB classifier learnt and trained on a same dataset and validated on the other two datasets independently (grey).



zak performed significantly better on cross-validation and Sartorelli shows the lowest differentiation between cross-validation and the independent test with almost the same average error rate on the cross-validation set compared to Cao. Although the differentiation of average error rate on the cross-validation set and independent test set is high in Tomczak, this model produced the best models in terms of the lowest overall error rate. This figure raises the idea that Tomczak is the most informative dataset since it can model any dataset, regardless of the gene selection method, significantly better than the other alternatives. This will be discussed in more detail in the *Extraction of infotmative genes* section.

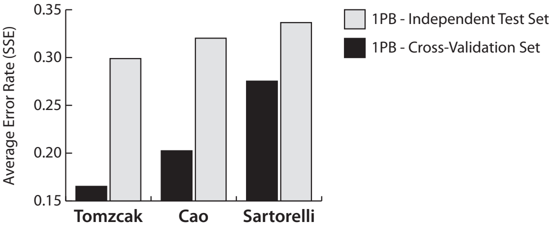
Comparison of gene selections with differing informativeness

We now look into how the different gene selections impact on the average error rate of the 1PB classifier for both cross-validation and the independent test. Figure 2 demonstrates the performance of the 1PB classifier in modeling datasets generated using different gene selections. Clearly, unlike Sartorelli, genes selected from Tomczak and Cao show very good performances on cross-validation. However, by looking at the average error rate of 1PB on independent test sets, we can see that the models learnt on Cao over-fitted the data and performed poorly on the independent test set (with the SSE of 0.32) whereas Sartorelli shows the lowest differentiation between the two sets. Overall the Tomczak selection performed the best both on cross-validation and the independent test.

It is important to adopt a methodology that can generate an accurate gene regulatory network, moreover, it is crucial to generate a model that can capture the significant genes and distinguish informative genes from uninformative ones. For this purpose, we added 50 randomly selected genes with high p-values (which imply less relatedness to Myogenesis) from the distribution. This also has the effect that it will increase the complexity of the datasets.

Figure 3 shows that there is a similar pattern on the average error rate of cross-validation. The additional random genes do not seem to affect Cao. It does, however, have an interesting impact on Sartorelli. The models learnt on Sartorelli (see **Additional file 1**) performed even poorer than SNB on the independent data sets and showed no significant changes when using different datasets for training. It is interesting because we know that the Sartorelli dataset is noisy and biologically complex and adding the random genes, which increases the complexity of the models in terms of more nodes and increases the risk of spurious links, produces a classifier which appears to be unable to capture the real gene interactions. The error rate and variance of models learnt on the Sartorelli selection is significantly high in comparison with Tomczak. By comparing figures 2 and 3, we can conclude that simpler and cleaner datasets tend to perform more reliably and have more stability while increasing the complexity. Since it is important to validate these models according to their variances, we demonstrated the average variance of each model on cross-validation and the independent test set in Additional file 1, Figure S8. Interestingly, we can see a similar pattern in the classifiers' variance in comparison with the average error rate (**Figure**

Figure 3 - The investigation of inference of adding more complexity to the model. We investigated the inference of adding more complexity to the model by adding 50 randomly selected genes as uninformative on 1PB classifier performance. In this figure we compare the average error rate of 1PB classifier after adding 50 uninformative genes to the model.



3). It is clear that we can raise the same conclusion as the simpler and cleaner datasets perform better than more noisy and complex ones. In this study, Tomczak performed favorably both in terms of bias and variance.

It is crucial to investigate if these findings are reproducible and are not prone to the number of samples and time points per dataset. Therefore, we applied our model on three synthetic datasets that have been generated by manipulating the biological, experimental, and model complexity of their known network structure using SynTReN application (Van den Bulcke et al., 2006). Additional file 1, Figure S9 illustrates that we can see a very similar pattern as we have seen on a real data where there is an increase on the average error rate of models learnt on multiple synthetic datasets with increasing biological variability. In the next section, before examining if these models can help us to capture the interactions in more complex datasets, we will investigate how well these models separate the informative genes from uninformative ones.

Extraction of informative genes

In order to test the ability of classifiers to separate informative genes from uninformative ones, we have looked at the result of the Kolmogorov-Smirnov test (KS test) on the ranking of genes according to their average error rate using a given model. Using this algorithm, we calculated the p-value, KS test, and the result of investigating the differentiation hypothesis along with the models' bias or variance. The results of this investigation are displayed in Additional file 1, Table S1 where Cao and Tomczak performed very well on cross-validation both in terms of bias and variance. However, models learnt on Sartorelli fail to separate between informative genes and uninformative genes as the scores are generally very low.

Generally, Tomczak outperformed Sartorelli and Cao and can be chosen as the most informative dataset in this study. Models learnt on Tomczak generated the lowest bias and variance and produced the best separation. In contrast, Sartorelli is the noisiest and less informative dataset while it failed to handle any increases in complexity (both biological and model wise) and generates models with highest bias and variance which also cause disability to separate informative genes from the others. Now the question is whether we can use a simpler and cleaner dataset to model more complex ones. In the next section we show how we tackled this question.

Analysis of the use of simpler dataset to model more complex one

In this section, we investigate the improvement or deterioration of genes selected by Tomczak on the Sartorelli dataset. Figure 4 shows the average improvement or deterioration of ranks of myogenesis-related genes, top 100 genes (most informative), and 50 randomly selected genes (uninformative) in Sartorelli. We compared the original rank of each gene (which can be any number between 1 and 150 derived from its p-value comparing to others) with its rank based upon the ability of a model trained on Tomczak to predict gene's value in Sartorelli. Moreover, we evaluate the improvement or deterioration of genes rankings in our model with the ones generated using

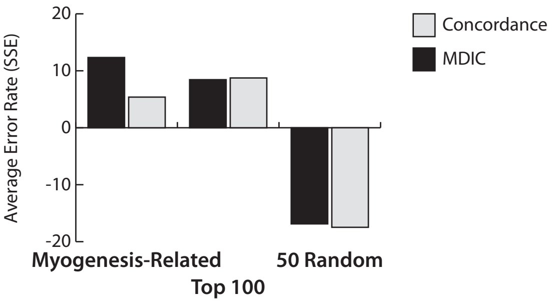


Figure 4 - The improvement or deterioration of genes ranking in Sartorelli. Firstly, we selected 100 informative and 50 uninformative genes using Tomczak dataset and extracted their ranks in Sartorelli. Secondly, we trained 1PB classifier on Tomczak and tested on Sartorelli. Finally, we ranked genes according to the average error rate of 1PB classifier in predicting their values in Sartorelli. This figure illustrates the average improvement or deterioration of Myogenesis-Related, Top 100, and 50 randomly selected genes in Sartorelli generated with our method and the gene rankings generated by concordance model.

the concordance model described by Lai *et al.* (2009). We can clearly see that the model learnt on Tomczak can capture the informative genes in Sartorelli and improve their rank whereas uninformative genes have been pushed down (almost 17 places in average) in the ranking by the classifier. Additionally, the improvement is even more pronounced for myogenesis-related genes with 12.33 places in average, which is significantly better than others with $P < 0.004$ generated using KS test, and as expected top 100 genes has been improved by 8.44 places. Even though both methods perform similarly on improving the ranks of top 100 and deteriorating the ranks of 50 randomly selected genes, the improvement of ranks for myogenesis-related genes are much more pronounced in our model than in the concordance model (improvement of 5.38 places).

Myh7 and *Tor3a* are two examples of significant improvements in Sartorelli dataset. *Myh7*, which originally ranked 101, improved 96 places to rank 5 (rank 55 in concordance model). During the learning phase it has been linked to four other genes of which three of them are myogenesis-related. These genes, in both datasets, have direct correlations and can represent each other in terms of prediction and validation. However, *Tor3a* has a very low rank in both dataset and yet improved 107 places from 128 to 21 (rank 31 in concordance model). It has been linked to *Prune* which also improved 106 places (from 131 to 25, 100 in concordance model). All three genes mentioned above have been selected as informative genes from Tomczak and yet placed into the bottom 50 due to the quality of Sartorelli dataset. These were some examples of the ability of model to pull out informative genes from a distribution (**Figures S10A** and **S10B**).

Although the overall improvement on myogenesis-related genes is significantly high, we were concerned why this model failed to improve the rank of some genes like *Id3* which dropped from rank 1 in Sartorelli to 133 (rank 51 in concordance model). In the learning process, *Id3* has been linked to 4 genes which are: *Fabp3*, *Rbm38*, *X99384*, and *Slco3a1*. Now in order to answer the question, firstly, we validate the relatedness of these genes to *Id3* in Tomczak dataset to investigate if they are significant and can represent *Id3*. Secondly, we study the expression level of these genes in Sartorelli to identify the reason why this model failed dramatically in predicting the *Id3* value.

Additional file 1, Figure S11 demonstrates the expression level of *Id3* along with its parent/children in both Tomczak and Sartorelli datasets. In Tomczak we can clearly see that there is an inverse relationship between *Id3* and the other 4 genes which is very significant. While the differentiation state changes, *Id3* drops from the expression level of approximately 11 to 8.5 and similarly its relatives show an increase of about 2 points in their expression values. This supports the assumption of the relatedness of these genes to *Id3* in the learning process on Tomczak dataset. However, considering that *Id3* is still very significant in Sartorelli, *Id3* parent/children show no variation and simply are not significant. As a conclusion, this model failed to predict *Id3* expression value and as a result the rank of *Id3* dropped 132 places most probably due to the quality and biological variation of Sartorelli dataset. Since we aim to overcome the lack of overlap on the gene regulatory network studies across species and platforms, the natural extension of the work in this paper would be to explore how this model can be used on datasets from multiple biological systems with increasing complexity. Moreover, it would be valuable to consider methods such as model averaging (Madigan and Raftery, 1994) that has been shown better generalization in classifier's accuracy. Consequently, it improves the performance of classifiers in identifying the most informative genes and avoids deterioration of cases like *Id3*. Furthermore, dynamic Bayesian networks can be adopted when learning from time-series data in order to handle auto-regulation and feedback loops, two key components of regulatory networks in biological data (Shen-Orr et al., 2002; Lee et al., 2002).

CONCLUSIONS

In this study, we have investigated a number of different Bayesian classifiers and datasets for identifying firstly, subsets of genes that are related to myogenesis and muscle differentiation, and secondly the use of cleaner and more informative datasets in modelling more biologically complex datasets. We have shown that an appropriate combination of simpler and more informative datasets produce very good results, whereas models learnt on genes selected from more complex datasets performed poorly. We concluded that simpler datasets can be used to model more complex ones and capture the interactions among genes. Moreover, we have described that highly predictive and consistent genes, from a pool of differentially expressed genes, across independent datasets are more likely to be fundamentally involved in the biological process under study. In three published datasets, we have demonstrated that these models can explain the myogenesis-related genes (genes of interest) significantly better than others ($P < 0.004$) since the improvement in their rankings is much more pronounced. These results imply that gene regulatory networks identified in simpler systems can be used to model more complex biological systems. In the example of muscle differentiation, a myogenesis-related gene network may be difficult to derive from in vivo experiments directly due to the presence of multiple cell types and inherently higher biological variation, but may become evident after initial training of the network on the cleaner in vitro experiments. In order to validate our approach, firstly, we evaluated our model on synthetic datasets and secondly we performed comparisons between our approach and the method of Lai *et al.* (2009) which we call concordance model. It is shown that our model performs comparably in improving the ranks of informative genes and deteriorating the ranks of uninformative ones, but that the improvement of ranks for myogenesis-related genes is much more pronounced whilst additionally modelling the interactions among genes. However, it is necessary to develop other statistical measures so that the model can be quantified to distinguish different degrees of complexities and platforms whilst handling the auto-regulation and feedback loops within the network.

Authors' contributions

SYA, PACH and AT contributed equally to methods development and drafting the paper. PACH provided the biological insight on the datasets related to muscle differentiation. AT designed the algorithms and SYA developed the codes. PACH and AT supervised the study. All authors analyzed the data, read and approved the final manuscript.

Acknowledgements

This work was funded in part by a Royal Society International Joint Projects Grant (2006/R3).

Reference List

Bockhorst,J., Craven,M., Page,D., Shavlik,J., and Glasner,J. (2003). A Bayesian network approach to operon prediction. Bioinformatics. *19*, 1227-1235.

Buntine,W.L. (1996). A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering *8*, 195-210.

Cao,Y., Kumar,R.M., Penn,B.H., Berkes,C.A., Kooperberg,C., Boyer,L.A., Young,R.A., and Tapscott,S.J. (2006). Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters. EMBO J. *25*, 502-511.

Chickering,D.M., Heckerman,D., and Meek,C. (2004). Large-sample learning of Bayesian networks is NP-Hard. Machine Learning Research *5*, 1287-1330.

Fielding,A.H. (2007). Introduction to classification. In Cluster and classification techniques for the Biosciences, Cambridge University Press), p. 86.

Friedman,N., Geiger,D., and Goldszmidt,M. (1997). Bayesian network classifiers. Machine Learning *29*, 131-163.

Friedman,N. and Koller,D. (2003). Being Bayesian about network Structure. A Bayesian approach to structure discovery in Bayesian networks. Machine Learning *50*, 95-125.

Friedman,N., Linial,M., Nachman,I., and Pe'er,D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601-620.

Grossman,D. and Domingos,P. (2004). Learning Bayesian network classifiers by maximizing conditional likelihood. Proceedings of the 21st International Conference on Machine Learning *68*, 46-54.

Haynes,B.C. and Brent,M.R. (2009). Benchmarking regulatory network reconstruction with GRENDEL. Bioinformatics *25*, 801-807.

Heckerman,D. (1998). A tutorial on learning with Bayesian networks. In Learning in graphical models, (Dordrecht: Kluwer Academic Publishers), p. 301.

Iezzi,S., Di,P.M., Serra,C., Caretti,G., Simone,C., Maklan,E., Minetti,G., Zhao,P., Hoffman,E.P., Puri,P.L., and Sartorelli,V. (2004). Deacetylase inhibitors increase muscle cell size by promoting myoblast recruitment and fusion through induction of follistatin. Dev. Cell *6*, 673-684.

Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U., and Speed,T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. *4*, 249-264.

Jelier,R., Schuemie,M.J., Veldhoven,A., Dorssers,L.C., Jenster,G., and Kors,J.A. (2008). Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. *9*, R96.

Kohavi, R. Wrapper for performance enhancement and oblivious decision graphs. 1995. Stanford University, Computer Science Department.

Ref Type: Thesis/Dissertation

Lai,Y., Eckenrode,S.E., and She,J.X. (2009). A statistical framework for integrating two microarray data sets in differential expression analysis. BMC. Bioinformatics. *10 Suppl 1*, S23.

Lam,W. and Bacchus,F. (1994). Learning Bayesian belief networks (an approach based on the MDL principle). Computational Intelligence *10*, 1-31.

Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I., Zeitlinger,J., Jennings,E.G., Murray,H.L., Gordon,D.B., Ren,B., Wyrick,J.J., Tagne,J.B., Volkert,T.L., Fraenkel,E., Gifford,D.K., and Young,R.A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Ma,H.W., Kumar,B., Ditges,U., Gunzer,F., Buer,J., and Zeng,A.P. (2004). An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. *32*, 6643-6649.

Mac Nally,R. (2000). Regression and model-building in conservation biology, biogeography and ecology: the distinction between - and reconciliation of - 'predictive' and 'explanatory' models. Biodiversity and Conservation *9*, 655-671.

Madigan,D. and Raftery,A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association *89*, 1535-1546.

Miron,M., Woody,O.Z., Marcil,A., Murie,C., Sladek,R., and Nadon,R. (2006). A methodology for global validation of microarray experiments. BMC. Bioinformatics *7*, 333.

Murphy,K.P. (2001). The Bayes Net toolbox for Matlab. Computing Science and Statistics: Proceedings of the Interface *33*, 331-350.

Pearl,J. (1986). Fusion, propagation, and structuring in belief networks. Artificial Intelligence *29*, 241-288.

Peña,J.M., Björkegren,J., and Tegnér,J. (2005). Learning dynamic Bayesian network models via cross-validation. Pattern

Recognition Letters *26*, 2295-2308.

Rhodes,D.R., Barrette,T.R., Rubin,M.A., Ghosh,D., and Chinnaiyan,A.M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res. *62*, 4427-4433.

Schwarz,G. (1978). Estimating the dimension of a model. The Annals of Statistics *6*, 461-464.

Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D., and Friedman,N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. *34*, 166-176.

Shen-Orr,S.S., Milo,R., Mangan,S., and Alon,U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. *31*, 64-68.

Stone,M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society *36*, 111-147.

Su,J. and Zhang,H. (2006). Full Bayesian network classifiers. Proceedings of the 23rd International Conference on Machine Learning *148*, 897-904.

Tobler,J.B., Molla,M.N., Nuwaysir,E.F., Green,R.D., and Shavlik,J.W. (2002). Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. Bioinformatics. *18 Suppl 1*, S164-S171.

Tomczak,K.K., Marinescu,V.D., Ramoni,M.F., Sanoudou,D., Montanaro,F., Han,M., Kunkel,L.M., Kohane,I.S., and Beggs,A.H. (2004). Expression profiling and identification of novel genes involved in myogenic differentiation. FASEB J. *18*, 403-405.

Van den Bulcke,T., Van,L.K., Naudts,B., van,R.P., Ma,H., Verschoren,A., De,M.B., and Marchal,K. (2006). SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC. Bioinformatics. *7*, 43.

Xu,X., Wang,L., and Ding,D. (2004). Learning module networks from genome-wide location and expression data. FEBS Lett. *578*, 297-304.

Zhang,C., Lu,X., and Zhang,X. (2006). Significance of gene ranking for classification of microarray samples. IEEE Transactions on Computational Biology and Bioinformatics *3*, 312-320.

APPENDIX

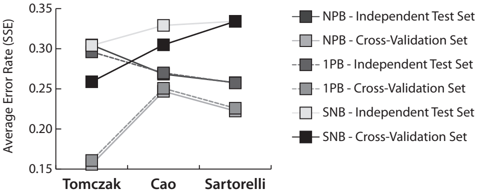


Figure S1 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Tomczak dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.

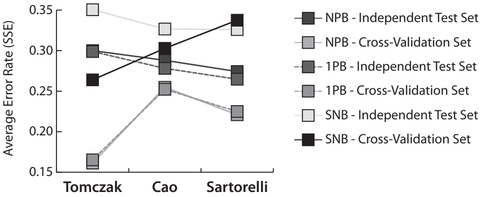


Figure S2 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Tomczak dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.

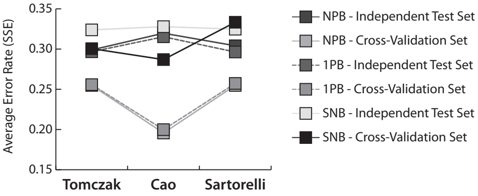


Figure S3 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Cao dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.

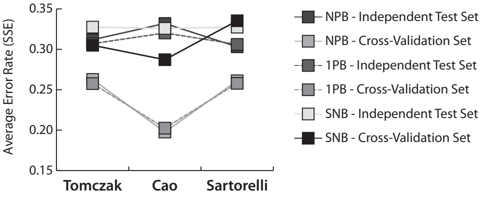


Figure S4 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Cao dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.

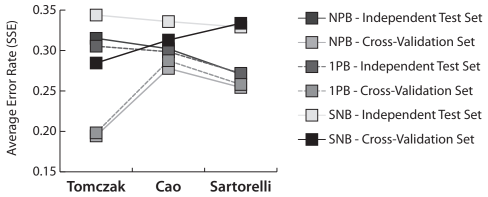


Figure S5 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Sartorelli dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.

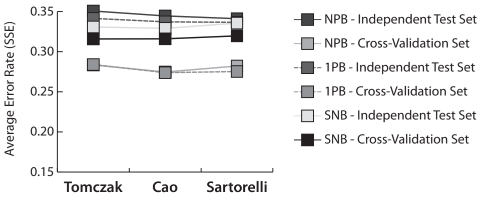


Figure S6 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Sartorelli dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.

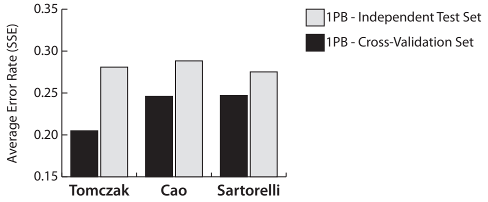


Figure S7 - The comparison of the differences between cross-validation set and independent test set on average error rates of 1PB classifier (extracted from figure 1).

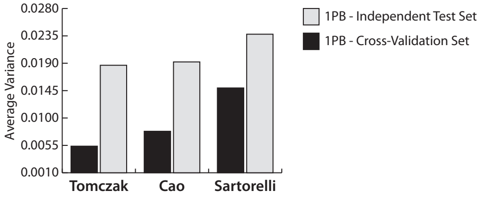


Figure S8 - The investigation of inference of adding more complexity to the model by adding 50 randomly selected genes as uninformative on 1PB classifier performance. In this figure we compare the average variance of 1PB classifier after adding 50 uninformative genes to the model.

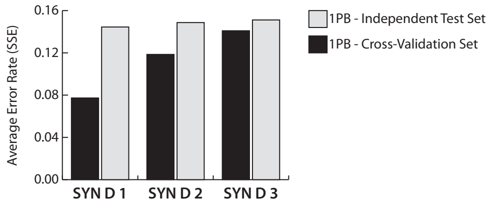


Figure S9 - This figure illustrates the performance of 1PB classifier on modeling three synthetic datasets generated using SynTReN application by manipulating the biological and experimental complexity. There is an increase of the biological variability on three datasets which matches an increase on the average error rate of models learnt.

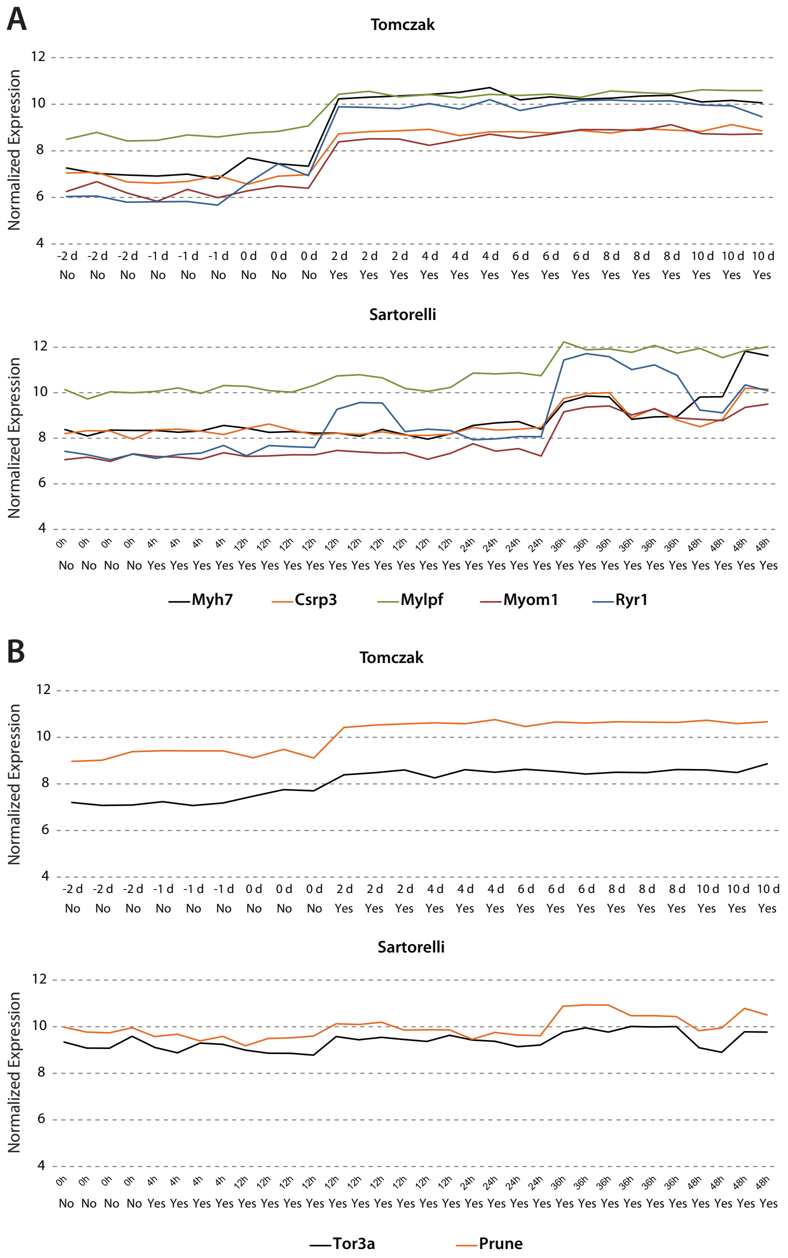


Figure S10 - A).The expression level of *Myh7* along with its parent/children in both Tomczak and Sartorelli datasets. In Tomczak we can clearly see that there is a strong relationship between *Myh7* and the other 4 genes. Moreover, in Sartorelli dataset the correlation still exists between *Myh7* and *Csrp3*, *Mylpf*, *Myom1*, and *Ryr1* even though it is not as strong as Tomczak. **B)** The expression level of *Tor3a* along with its parent in both Tomczak and Sartorelli datasets. In Tomczak we can clearly see that there is a good relationship between *Tor3a* and *Prune*. Moreover, in Sartorelli dataset the correlation still exists between *Tor3a* and *Prune*. This figure is an example of a large improvement of rank of a given gene after training on Tomczak. The x-axis represents both the time points and the differentiation status.

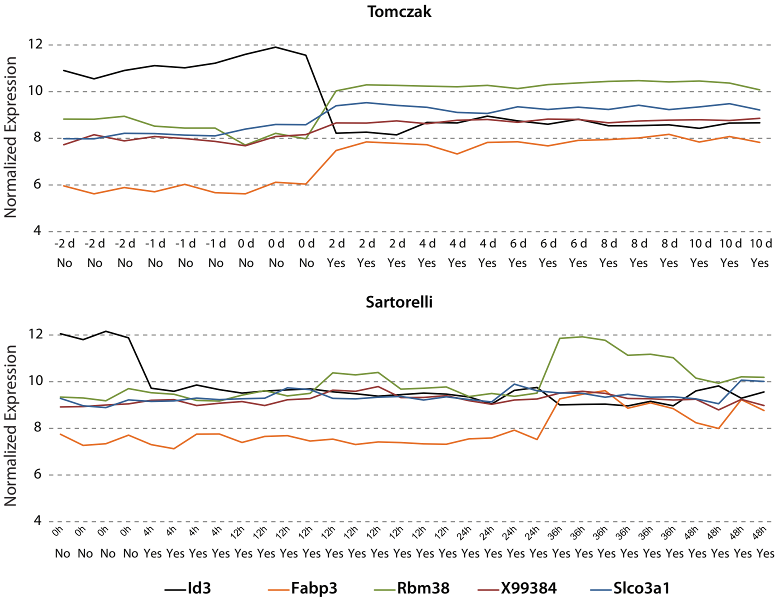


Figure S11 - The expression level of *Id3* along with its parent/children in both Tomczak and Sartorelli datasets. In Tomczak we can clearly see that there is an inverse relationship between *Id3* and the other 4 genes while Sartorelli dataset shows no significant correlations between *Id3* and *Fabp3*, *Rbm38*, *X99384*, and *Slco3a1*. This figure is an example of a large deterioration of rank of a given gene after training on Tomczak. The x-axis represents both the time points and the differentiation status.

Table S1 - Differentiation Hypothesis. Investigating how well the models can separate the informative and uninformative genes from each other. Firstly, we ranked genes according to their average error rate and variance. Secondly, using Kolmogorov-Smirnov test and original ranking list, we explored which model can separate the informative genes from uninformative genes the best.

		Error Rate (SSE)		**Variance**	
Gene Selection		Cross-Validation Set	Independent Test Set	Cross-Validation Set	Independent Test Set
» Tomczak	Differentiation Hypothesis	TRUE	TRUE	TRUE	TRUE
	P-value	5.02E-24	9.77E-10	5.02E-24	3.68E-05
	Kolmogorov-Smirnov Test	0.880198	0.552871	0.880198	0.394257
	Average Performance	0.165259	0.298921	0.00537	0.018667
Cao	Differentiation Hypothesis	TRUE	TRUE	TRUE	TRUE
	P-value	1.89E-22	6.16E-06	1.91E-20	0.004314
	Kolmogorov-Smirnov Test	0.850297	0.425347	0.810693	0.295842
	Average Performance	0.202472	0.320211	0.007819	0.019219
Sartorelli	Differentiation Hypothesis	FALSE	TRUE	FALSE	FALSE
	P-value	0.443901	0.007507	0.527435	0.104457
	Kolmogorov-Smirnov Test	0.145941	0.282178	0.136832	0.205149
	Average Performance	0.275287	0.336551	0.014939	0.023772

Table S2 - The specification of three synthetic datasets generated for the purpose of the validation and reproduction of the result of applying our model on real microarray datasets used for this study. Three datasets have been generated on the well-described network structure of *E. coli* (Ma *et al.*, 2004) which contains 1330 number of nodes and 2724 interactions. Average performance is measured based on SSE/Variance.

	SYN D 1	**SYN D 2**	**SYN D 3**
Burnin point	2000	2000	2000
Number of Experiments	15	15	15
Number of Samples per experiment	2	2	2
Number of Nodes	1000	1000	1000
Number of Background nodes	0	0	0
Probability for complex 2-regulator interactions	0.3	0.5	0.7
Biological noise	0.1	0.3	0.5
Experimental noise	0.1	0.3	0.5
Noise on correlated inputs	0.1	0.3	0.5
Number of External nodes	0	0	0
Number of Correlated external nodes	0	0	0
Sub network selection method	Cluster Addition		
Random seed	13	13	13

Interspecies translation of disease networks increases robustness and predictive accuracy

Seyed Yahya Anvar[1,*], Allan Tucker[2], Veronica Vinciotti[2], Andrea Venema[1], Gert-Jan B. van Ommen[1], Silvère M. van der Maarel[1], Vered Raz[1] and Peter A.C. 't Hoen[1]

Gene regulatory networks give important insights into the mechanisms underlying physiology and pathophysiology. The derivation of gene regulatory networks from high-throughput expression data via machine learning strategies is problematic as the reliability of these models is often compromised by limited and highly variable samples, heterogeneity in transcript isoforms, noise, and other artifacts. Here, we develop a novel algorithm, dubbed Dandelion, in which we construct and train intraspecies Bayesian networks that are translated and assessed on independent test sets from other species in a reiterative procedure. The interspecies disease networks are subjected to multi-layers of analysis and evaluation, leading to the identification of the most consistent relationships within the network structure. In this study, we demonstrate the performance of our algorithms on datasets from animal models of oculopharyngeal muscular dystrophy (OPMD) and patient materials. We show that the interspecies network of genes coding for the proteasome provide highly accurate predictions on gene expression levels and disease phenotype. Moreover, the cross-species translation increases the stability and robustness of these networks. Unlike existing modeling approaches, our algorithms do not require assumptions on notoriously difficult one-to-one mapping of protein orthologues or alternative transcripts and can deal with missing data. We show that the identified key components of the OPMD disease network can be confirmed in an unseen and independent disease model. This study presents a state-of-the-art strategy in constructing interspecies disease networks that provide crucial information on regulatory relationships among genes, leading to better understanding of the disease molecular mechanisms.

---

1 Center for Human and Clinical Genetics, Leiden University Medical Center, the Netherlands. **2** Center for Intelligent Data Analysis, School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK.

* To whom correspondence should be addressed at: s.y.anvar@lumc.nl

PLoS Computational Biology, 2011,
Nov 3; 7:11; doi:10.1371/journal.pcbi.1002258

AUTHOR SUMMARY

The identification of gene regulatory networks can provide vital information on biological processes. Despite numerous advancements in developing machine learning strategies, the stochastic nature of such biological systems complicates the construction of robust and reliable network structures. In recent years, the use of cross-species datasets enabled scientists to better understand the molecular mechanisms that are associated with human disorders. However, it also presents a challenge in dealing with especially difficult mapping of protein orthologues, alternative transcript splicing, noise, or other artifacts. Here, we developed a novel algorithm for constructing interspecies disease networks that provide accurate predictive value over the disease phenotype and gene expression. We show that the disease-association of potential key regulators that play a role in interspecies disease networks can be reproduced and validated in an unseen and independent model system. This study presents a novel strategy for constructing networks that can be translated across species whilst providing a comprehensive view of regulatory relationships associated with the disease.

INTRODUCTION

The degree to which gene products appear in the cell and exert their function is regulated through interactions with other genes. This interconnectivity implies that the identification of gene regulatory networks is vital for understanding the phenotypic impacts of gene defects and the associated complications (Schadt, 2009; Goldstein, 2009; Karlebach and Shamir, 2008; Barabasi et al., 2011). The dawn of high-throughput technologies such as genome-wide sequencing and microarray experiments has increased our understanding of molecular behavior at the transcriptional level. Although these large-scale datasets provide crucial information about both the presence and relative abundance of RNA transcripts, they also introduce an important challenge in providing a comprehensive view of molecular mechanisms and regulatory relationships among genes with different underlying phenotypic conditions.

The presence of this obstacle calls for developing robust machine learning models that can be used for generating gene networks in which their transcriptional changes can affect phenotypic outcome. However, building a network that involves thousands of genes and millions of interactions is extremely problematic and requires a great quantity of experimental data for the valid interpretation of biological causes for a given phenotype. Furthermore, the validity of gene regulatory networks is often affected by limited and highly variable samples, heterogeneity in transcript isoforms, noise and other artifacts (Raj and van Oudenaarden, 2008; Kluger et al., 2003; Shahrezaei and Swain, 2008; Pedraza and van Oudenaarden, 2005). Therefore, a probabilistic approach is needed to identify and predict interconnected transcriptional behaviors that give rise to disease outcome (Pache et al., 2008) and to, ultimately, offer potential targets for therapeutic intervention and drug development. Among the possible statistical models, Bayesian networks have been an important concept for modeling uncertain systems (Pearl, 1988; Friedman, 2004; Friedman et al., 2000; Segal et al., 2003). Bayesian networks can represent complex stochastic relationships between genes and are capable of integrating different types of data (i.e. phenotype and genotype categorical information as well as gene expression data). In addition, the probabilistic nature of such networks can accommodate noise and missing data by weighting each information source according to its reliability. In contrast to many statistical models, the transparent nature of Bayesian networks (in terms of the graphical structure and local probability distributions) leads to better interpretation and understanding of the underlying biological regulation of the disease.

The high dimensionality of the genome wide expression profiling datasets and the limited number of available samples complicates the derivation of robust network structures. Methods such as the use of prior knowledge about biological interactions (Segal et al., 2003; Pe'er et al., 2002; Steele et al., 2009) have been shown to successfully reduce the search space and to make networks more robust. This method works for well-studied diseases or biological systems, but is not likely to identify novel regulatory interactions underlying the molecular mechanisms of rare or complex disorders. In addition, this bias can falsely expose the network to sample differences in the absence of a disease-related biological cause. In this study, we hypothesize that biologically relevant relationships between genes are often conserved across species. Thus, the robustness and stability of a gene network should increase when modeling regulatory networks using related datasets from different species. Moreover, we hypothesize that the relationships identified in an interspecies gene network should be biologically more meaningful. On the other hand, cross-species translation of networks is far from trivial given our limited knowledge of true protein orthologues and transcript variants coding for proteins with similar functions in different species. Therefore, we explore the performance of a novel algorithm that combines our previously published model for learning regulatory interactions from multiple datasets of increasing complexity (Anvar et al., 2010) with an interspecies translation and validation regime, named *Dandelion algorithm*. We show that the supplementation of this algorithm with a modeling-driven selection of transcripts coding for orthologous proteins (*exhaustive* Dandelion algorithm) significantly improves the *robustness* and stability of the *interspecies network*, when compared to a standard approach in which expression levels of different transcripts for the same gene are summarized (*naïve* Dandelion algorithm). We also show that the potential regulatory relationships that play a role in *interspecies disease networks* can be reproduced and validated in an unseen and independent model system.

In this study, three publicly available microarray datasets from *Drosophila* (Chartier et al., 2009), mouse (Trollet et al., 2010), and human (Anvar et al., 2011) that are all concerned with oculopharyngeal muscular dystrophy (OPMD) have been chosen to gain insight into the key regulators of the disease. These datasets are described in Table 1. OPMD is a late-onset progressive muscular disorder for which the underlying molecular mechanisms are largely unknown. This autosomal dominant muscular disorder has an estimated prevalence of 1 in 100,000 worldwide (Fan and Rouleau, 2003). OPMD is caused by the expansion mutation of a homopolymeric alanine stretch at the N-terminus of the Poly(A) Binding Protein Nuclear 1 (PABPN1) by 2-7 additional Ala residues (Brais et al., 1998). Although PABPN1 is ubiquitously expressed, the clinical and pathological features of OPMD are restricted to a subset of skeletal muscles, causing progressive *ptosis*, *dysphagia*, and limb muscle weakness. *Drosophila* and mouse models with muscle-specific overexpression of expanded PABPN1 recapitulate progressive muscle weakness in OPMD (Chartier et al., 2006; Davies et al., 2005). However, the potential artifact, heterogeneity in transcript isoforms, and the presence of overexpression side-effects in OPMD animal models and limited patient materials complicate the identification of key regulators of OPMD. With the analysis of these datasets, we demonstrate that modeling of *interspecies disease networks* increases the *robustness* of the networks and aids in the identification of key regulators of the disease.

METHODS

Model of Interspecies Networks using Dandelion Algorithm

To construct *interspecies networks* that can accurately predict the disease phenotype and provide a comprehensive view of molecular relationships that underlie the disease-associated biological processes, we developed a novel *Dandelion algorithm* with multi-layers of analysis and evaluation criteria. A schematic presentation of this approach can be found in Figure 1. In addition, the definition of nomenclatures (italicized terms) used in this study is provided in the Table S1 in Text S1.

The procedure starts with the identification of the disease-associated modules by assessing the association of transcriptional profiles with the disease state. In this study, gene modules are defined according to current KEGG (Kyoto Encyclopedia of Genes and Genomes) annotation of molecular pathways to ensure functional relationships among genes within the same cluster. After identification of the *disease module*, the set of genes in the *disease module* is supplemented with a set of randomly selected genes for the purpose of network performance estimation and evaluation. The Dandelion algorithm integrates three recurring phases of training and independent testing with the use of multiple datasets derived from the different biological systems. This involves a reiterative selection of one species as an organism in which *intraspecies* gene regulatory networks are constructed. Cross-validation is used for learning and optimization of the intraspecies network structure. Some partitions were purely used for testing the intraspecies network to ensure, in all experiments, that the test data is previously unseen. Datasets from the other species are used for *interspecies* translation, independent testing and validation of the constructed disease networks. The construction of intraspecies Bayesian networks is governed by our previously published optimization procedure (Anvar et al., 2010). To ensure that these *interspecies networks* are derived from a disease-related biological cause, the *specificity* and *sensitivity* of the networks for prediction of the disease phenotype are assessed. Moreover, the *robustness* and *translatability* at different *confidence* thresholds are evaluated. After defining the *interspecies disease domains*, a subset of genes is selected for unbiased examination of reproducibility and validity of disease-related transcriptional changes in an unseen and independent model system. The detailed outline of the procedure, depicted in Figure 1, is provided in the following subsections.

Disease Modules. Disease modules have been identified according to our previously published study (Anvar et al., 2011) in which we performed an integrated transcriptome analysis to identify the most significant molecular pathways that are associated with the OPMD across species.

Bayesian Network Structure Learning. A Bayesian network encodes the joint probability distribution of a set of random variables. It consists of a directed acyclic graph (DAG) that represents conditional independencies between variables, and conditional distributions at each node in the graph. Bayesian network classifiers are a special case of Bayesian networks where one node represents some discrete class to be predicted. Here, each node in the graph represents a gene transcript (or gene) and the class node represents the disease states. In order to learn the Bayesian network structure of a gene network, the algorithm approximates the likely graphical model by searching the space of possible networks via single-arc changes that improves some score. We use a simulated annealing search in conjunction with the Bayes Information Criterion (BIC) as a scoring metric (Schwarz, 1978). Simulated annealing performs competitively with other optimization methods as it aims to avoid local maxima (Friedman et al., 1997). There is a trade-off between simplicity of model with one that can accurately identify the empirical distribution of gene expression profiles and predict the disease phenotypic outcome. For this reason the BIC is used as it is less prone to overfitting through the use of a penalizing term for overly complex models.

Table 1 – Overview of microarray datasets and networks constructed by Dandelion algorithm.

					Cross-Validation	Number of Networks		
Species	**Tissue**	**Samples**	**Age / Time-Point**	**GEO Accession**	**Number of folds**	**Human**	**Mouse**	**Drosophila**
Human	Quadriceps	4 Symptomatic 18 Controls	49 – 60 Year-old 17 – 89 Year-old	GSE26605	4	4	24	24
Mouse	Quadriceps	17 OPMD 16 Wild-type	6, 18, 26 week-old per genotype	GSE26604	6	24	6	36
Drosophila	Adult thoracic muscles	18 OPMD 18 Wild-type	1, 6, 11 day-old per genotype	-	6	24	36	6

The initial state of the structure is an empty DAG with no links. In order to alter the network structures, three operations have been used within the simulated annealing procedure. These operators are *adding*, *removing*, or *swapping* links to generate a new network which can be either accepted or rejected based on its overall score and the current temperature. The outline of this algorithm can be found in the Protocol S1 in Text S1.

In this study, the initial temperature (t_0) has been set to 10 and it terminates at 0.001 (t_n), according to our previously published optimization procedure (Anvar et al., 2010). The number of iterations (*maxfc*) has been set to 1000 in respect to the number of nodes available in the network. The training dataset is described as *D*. For the training phase, the *mode* variable is set to "train" and the variable *networkMap* is set to empty. During the interspecies translation and testing, the variable *mode* is set to "test" and the variable *networkMap* holds information on the regulatory relationships that are present in the network map constructed on training organism.

Construction of Interspecies Networks. The Dandelion algorithm takes multiple datasets from different species as input. In this study, we launch two classes of Dandelion algorithm. Firstly, the naïve Dandelion algorithm, where the expression patterns of gene transcripts are summarized by averaging the expression profiles of gene probes, to provide one expression profile per gene. This enables direct mapping of expression profiles of orthologous genes when translating networks across species. This approach significantly simplifies the process of constructing network structures. Secondly, we developed the exhaustive Dandelion algorithm to overcome the limitations caused by heterogeneity in transcript isoforms, differences in annotation between organisms and technical factors (i.e. different microarray platforms). In the exhaustive algorithm, transcripts that are most likely to be coding for orthologous proteins are selected automatically in the modeling phase.

The procedure involves reiterative selection of one species for construction of the Bayesian network while other species are left aside for independent testing and validation of learnt disease networks. The highest-scoring intraspecies network structure is learnt according to the algorithm described in the Protocol S1 in Text S1. Before interspecies translation, in the exhaustive Dandelion algorithm, a detailed interaction map of a candidate intraspecies disease network of gene transcripts needs to be transformed to a network map of gene-gene relationships. This step can be omitted in the naïve Dandelion algorithm as the constructed intraspecies networks are already at the gene level.

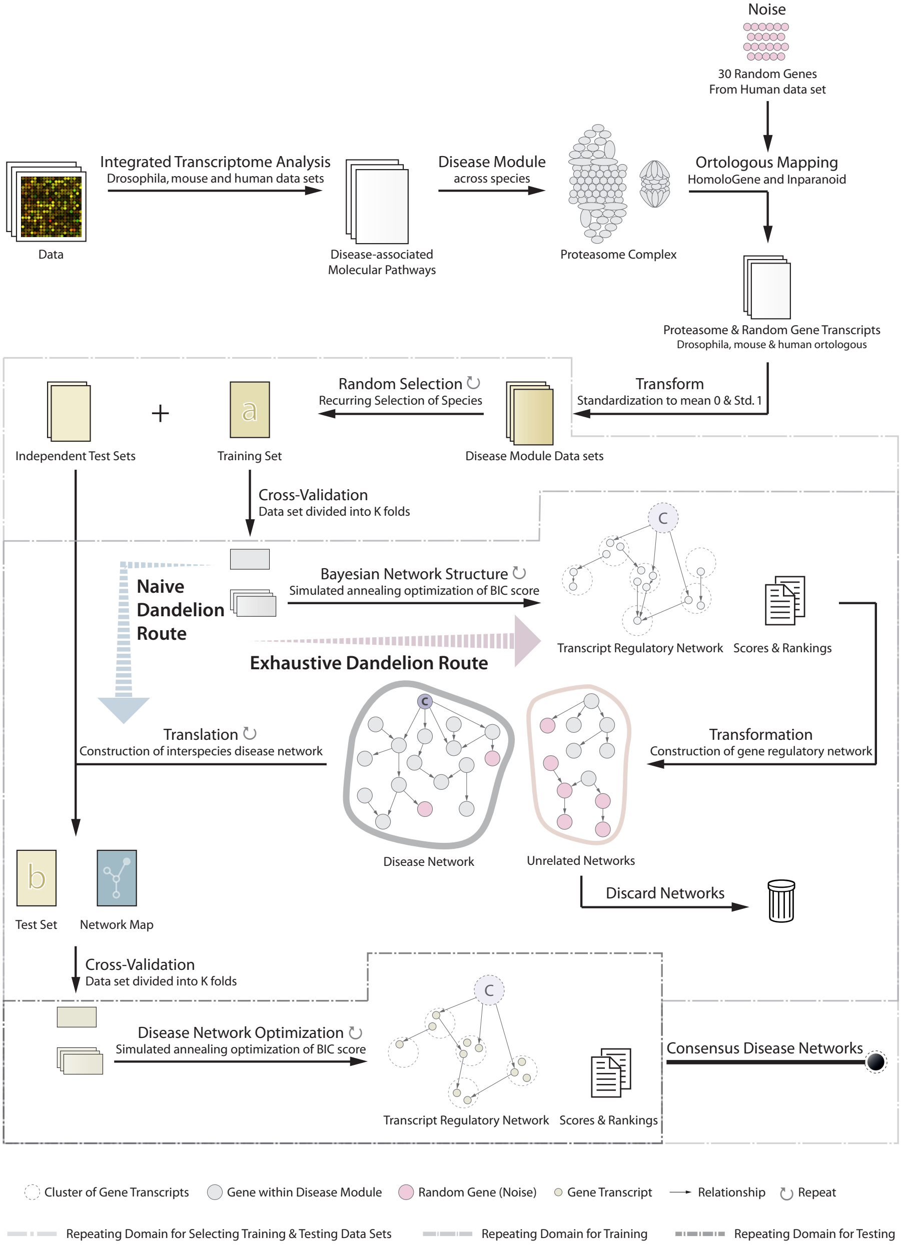

Noise
30 Random Genes
From Human data set
Data
Integrated Transcriptome Analysis
Drosophila, mouse and human data sets
Disease-associated
Molecular Pathways
Disease Module
across species
Proteasome Complex
Ortologous Mapping
HomoloGene and Inparanoid
Proteasome & Random Gene Transcripts
Drosophila, mouse & human ortologous
Independent Test Sets
+
a
Training Set
Random Selection
Recurring Selection of Species
Disease Module Data sets
Transform
Standardization to mean 0 & Std. 1
Cross-Validation
Data set divided into K folds
Naive
Dandelion
Route
Bayesian Network Structure
Simulated annealing optimization of BIC score
Exhaustive Dandelion Route
C
Transcript Regulatory Network
Scores & Rankings
Translation
Construction of interspecies disease network
C
Disease Network
Unrelated Networks
Transformation
Construction of gene regulatory network
Discard Networks
b
Test Set
Network Map
Cross-Validation
Data set divided into K folds
Disease Network Optimization
Simulated annealing optimization of BIC score
C
Transcript Regulatory Network
Scores & Rankings
Consensus Disease Networks
Cluster of Gene Transcripts
Gene within Disease Module
Random Gene (Noise)
Gene Transcript
Relationship
Repeat
Repeating Domain for Selecting Training & Testing Data Sets
Repeating Domain for Training
Repeating Domain for Testing

Figure 1 – Schematic overview of the Dandelion algorithm for disease network analysis. The Dandelion algorithm involves three recurring stages of training and independent testing regime with the use of multiple datasets derived from different species. In the first step, disease modules are defined as the most consistently disease-associated molecular pathway across species. The disease module is supplemented by a set of randomly selected genes to assess the performance of the algorithm and to check for overfitting. These datasets are standardized to mean 0 and standard deviation of 1 across genes. The next step involves re-iterative selection of one species as an organism in which the gene regulatory network is constructed while others are left aside for independent testing and validation of learnt disease networks. For an intraspecies construction of disease network, dataset is divided into k-folds, using cross-validation, and regulatory relationships between gene transcripts are learnt using Bayesian network methodology enhanced by simulated annealing optimization of network BIC score. After applying confidence thresholds on relationship between genes, the disease network can then be translated to the expected interspecies disease network which we call a network map. Using the cross-validation and network optimization procedure the algorithm searches through the relationships found in the training dataset to find the best fit for interspecies representation of the disease network. These networks are then integrated by removing all the links with low confidence score across species.

Using the cross-validation and network optimization procedure, the algorithm searches through the relationships present in the network map (constructed on the training set) to find the best fit for the interspecies representation of the disease network. These networks are then integrated by removing all the links with a low confidence score to construct the consensus interspecies disease networks. The full algorithm details are outlined in the Protocol S2 in Text S1 where $Species_{train}$ and $train_{folds}$ represent the training dataset and the folding arrangements for the selected organism. Furthermore, the series of $Species_{test\ 1}$ … $Species_{test\ M}$ and $test_{folds\ 1}$ … $test_{folds\ M}$ represent the datasets and folding arrangements of organisms that are selected for independent test and validation. The logical variable exhaustive indicates the class of Dandelion algorithm (naïve in case of *false* and exhaustive in case of *true*) that needs to be performed. In this study, the human dataset is divided into 4 folds due to the limited number of patient samples. Mouse and *Drosophila* datasets are divided into 6 folds. The average *Sum of Squared Error* (SSE) and standard deviation (STD) are calculated for all nodes over these folds by predicting the measured expression values of genes (or gene transcripts) given the measurements taken from others. For the class node, the state of the disease is predicted given the expression profiles for genes (or gene transcripts) within the network structure. The number of iterations was set to 1000 for the training phase and was reduced to 500 during the interspecies translation of disease networks. The code is implemented in Matlab 2008b using the Bayes Net toolbox (Murphy, 2001).

Network Analysis and Evaluation. The proposed approach consists of three layers of analysis and evaluation. The constructed interspecies disease networks are assessed for their predictive accuracy towards the disease phenotype (class node) by calculation of the level of sensitivity and specificity. Furthermore, the Bayesian networks Sum of Squared Error (SSE) is calculated for prediction of the expression of all genes (or gene transcripts). Moreover, the level of robustness and translatability of the generated networks are evaluated. The stability and robustness of relationships between genes within the disease module are compared to those of the random genes at different confidence score thresholds. Confidence scores are the ratio of the number of times a link is found in the interspecies disease networks to the maximum number of times the link can possibly be found (based upon the number of folds). For approximating the level of translatability, the total number of links found during the training phase is compared to the number of links that were successfully translated to other species. Finally, the interspecies disease domains are defined based on the Markov blanket principle for the extension of the class node connectivity. In addition, unstable gene interactions are removed through assessment of the level of confidence in the relationships between genes. The interspecies disease domains are used to select a subset of genes to further study the reproducibility and validity of the observed relationships towards their association with the disease phenotype in an unseen and independent OPMD model system.

To assess the specificity of genes encoding for the proteasomal proteins in accurately predicting the disease states, we generated three additional gene sets. A set of 100 randomly selected genes, 87 genes within the ribosome pathway, and 70 randomly selected genes with the constraint of none being deregulated (ND) constitute the three genes sets that are used in a comparative analysis. The human dataset is used for cross-validation whilst mouse and *Drosophila* datasets were used for independent assessment of the constructed networks. Networks are evaluated on their sensitivity, specificity, and predictive accuracy towards the disease state (OPMD or control).

Microarray Datasets

The human, mouse, and *Drosophila* microarray datasets have been previously published (Chartier et al., 2009; Anvar et al., 2011; Trollet et al., 2010). The human and mouse datasets are publicly available at GEO repository under the accession numbers GSE26605 and GSE26604, respectively. In all datasets genome-wide expression profiles of skeletal muscles from OPMD are compared to controls. In case there are multiple probes for the same gene on the microarray platforms, these probes usually measure the expression levels of different transcripts from the same gene. The class node reflects the disease phenotype (control or OPMD) of each sample. A detailed description of these datasets can be found in Table 1.

Data Processing and Statistical Analysis

Microarray measurements were normalized using the quantile method. In addition, these datasets were standardized to mean 0 and standard deviation 1 across the genes. For the scope of this paper, the human proteasome-encoding genes were annotated using illuminaHumanv3BeadID package in R and the mouse and *Drosophila* homologous were annotated using HomoloGene and Inparanoid (*http://ncbi.nlm.nih.gov/homologene* and *http://inparanoid.sbc.su.se*, respectively) online databases. Previously published data were used to identify deregulated genes per species (Anvar et al., 2011). For cross-validation (Stone, 1974; Fielding, 2007) human data were divided into 4 folds (given the limited number of OPMD samples), while the other datasets were divided into 6 folds (**Table 1**). Human, mouse, and *Drosophila* datasets hold 108, 96, and 78 transcripts, respectively, which encode for 74, 56, and 53 genes (including genes encoding for the proteasome and a set of 30 randomly selected genes). The differences are due to limitations of mapping homologous genes or unavailability of expression data for certain genes in a particular species. The gene lists are provided in the Table S2 in Text S1.

Cell Model

IM2 cells stably transfected with normal (WTA) or expanded PABPN1 (D7E) and were compared to assess the predictive value of the interspecies modeling approach on an unseen OPMD disease model (Raz et al., 2011). Exogenous PABPN1 expression is under control of the desmin promoter. IM2 cells were proliferated in DMEM supplemented with 20% fetal calf serum, 0.5% chicken embryo extract, 5U/ml interferon gamma, at 33C and 10% CO2. Myotube fusion was induced by culturing in DMEM supplemented with 5% horse serum at 37C and 5% CO2 for four days, after which RNA was extracted from three independent cultures.

Quantitative RT-PCR Analysis

Total RNA was extracted using the TRIZOL reagent (Invitrogen) according to manufacturer's instruction. First strand cDNA was synthesized with random hexamer oligonucleotides and MMLV reverse transcriptase (First Strand Kit; Fermentas, according to manufacturer's instruction). 3.6ng cDNA was used per quantitative PCR reaction. qPCR was performed with SYBR green mix buffer (BioRad) and 7.5 pmole (per reaction) of forward and reverse primers in a 15 µL reaction volume. PCR conditions were as follows: 4 min at 95 °C followed by 40 cycles of 10 sec at 95 °C and 60 sec at 60 °C. The program was ended with 1 min at 60 °C. For each primer set, the specificity of the PCR products was determined by melting curve analysis. Expression levels were calculated according to the ΔΔCT method normalized to mHrpt, Desmin, and IM2 parental cells. The statistical significance was determined with the student's t-test. The list of primers used in this study is provided in the Table S3 in Text S1.

RESULTS

Identification of Disease Module

Previously we identified that the deregulation of the ubiquitin-proteasome system (UPS) is the predominant molecular pathway affected in OPMD animal models and patients (Anvar et al., 2011). The UPS, a cellular regulator of homeostasis, is highly dynamic machinery that involves protein ubiquitination and degradation steps. From the six UPS components, we found that only E3-ligases, deubiquitinating enzymes, and proteasome components are consistently and prominently deregulated in OPMD across species (Anvar et al., 2011). The proteasome is composed of core and regulatory subunits. We observed a substantial deregulation of proteasome and cytokine-induced proteasome (also known as immunoproteasome) encoding genes across species (**Figure 2**). To obtain more insight in the key components in the proteasome machinery that are aberrantly expressed in OPMD across species, we generated gene regulatory networks. Unique to the current approach, the networks were learnt on one species and evaluated on datasets from other species. This was done to only retain those links between genes that can be found across multiple species and that are more likely to be directly connected to the disease phenotype than links that are only found in a single species. For the interspecies translation we used two version of our newly developed Dandelion algorithm. The naïve variant is a straw man approach, where expression values for different transcripts of the same gene are first summarized. This approach was then further refined in the exhaustive Dandelion algorithm, where the model chooses the transcript that is most predictive for the expression value of a transcript in another species.

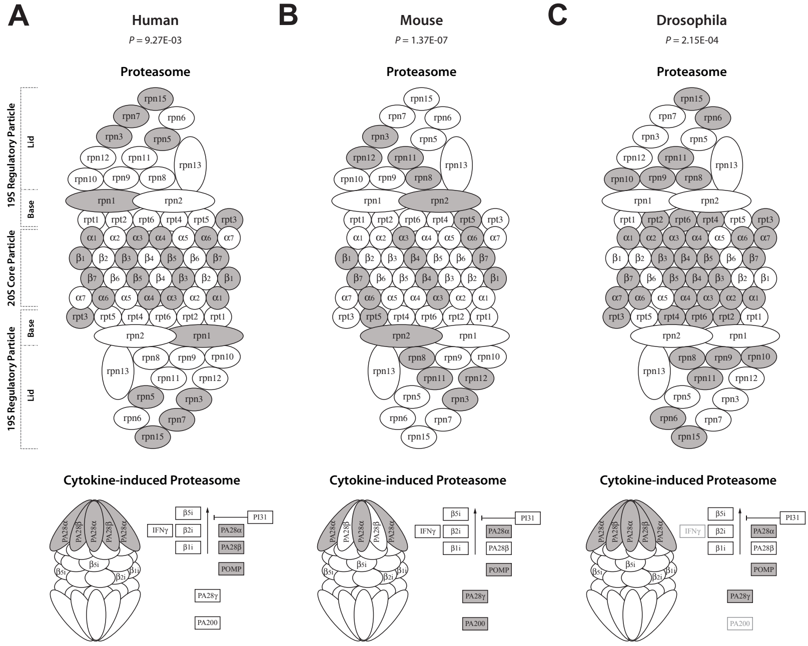


Figure 2 – OPMD-deregulation across different subunits of the proteasome in different species. There are widespread differences in gene expression (depicted in dark colors) between OPMD and control in the different functional subunits of proteasome and immunoproteasome in human (**A**), mouse (**B**) and *Drosophila* (**C**). The Significance of the association between the disease outcome and expression profiles of genes encoding for proteasome and immunoproteasome were previously calculated (Anvar et al., 2011) using the global test (Goeman et al., 2004).

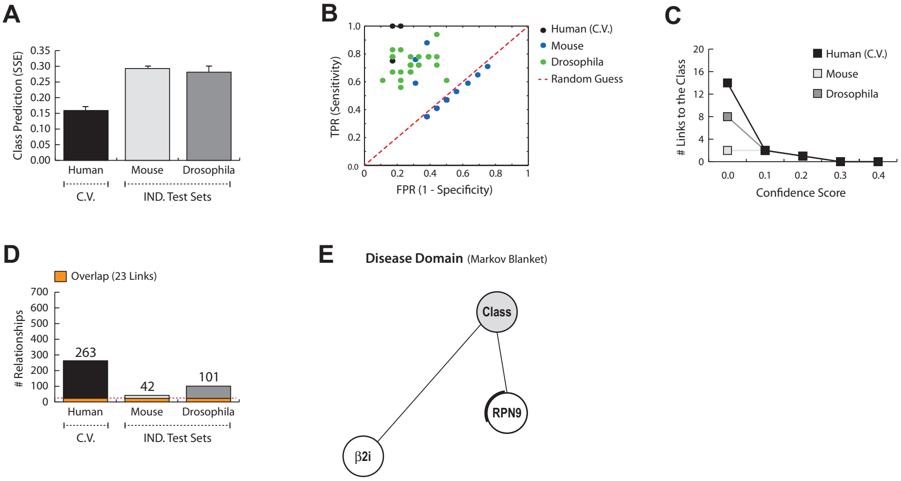


Figure 3 – Performance of the naïve Dandelion algorithm on constructing disease networks that are learnt on human and evaluated on human, mouse and *Drosophila* datasets. A) The average Sum of Squared Error (SSE) for prediction of the disease phenotype (OPMD vs. control) given the gene expression profiles within the disease networks learnt on human. The cross-validation set which is used during the training phase is depicted by *C.V.* and the independent test sets are grouped as *IND. Test Sets*. **B)** ROC space demonstrates the relative sensitivity and specificity of the generated networks in predicting the disease phenotype. The results from random expectations are illustrated by the red dash-line. **C)** Number of relationships between genes and the class node, after applying confidence thresholds, are depicted in line per species. **D)** The number of links found after interspecies translation and optimization of the disease networks within each species. The orange section, separated by red dash-line, represents the number of links that can be found in all species with the confidence threshold of 0.1. **E)** The interspecies disease domain is generated according to the Markov blanket criteria, after applying the confidence threshold of 0.1.

Naïve Construction of Disease Network

The process of constructing disease networks using naïve Dandelion algorithm initially starts by averaging the expression profiles of different gene transcripts in the human datasets. The summarized gene expression values were then used for the learning of intraspecies gene networks which consequently were translated to the other species. The interspecies networks were assessed for their predictive accuracy, sensitivity and specificity (**Figure 3**). The constructed interspecies networks predict the disease status (control vs. OPMD) of the unseen *Drosophila* and mouse samples with a moderate accuracy of 71% and 72%, respectively (**Figure 3A**). However, a large number of networks perform worse than random expectations, as evident from the ROC space (**Figure 3B**). This result indicates an overall low level of sensitivity and specificity in predicting the disease phenotype. Moreover, the networks are weak and unstable as they exhibit a very low level of translatability (**Figure 3C**). The low level of robustness, stability and translatability is also evident from the low percentage (8.7%) of relationships with the confidence score of ≥ 0.1 in the intraspecies networks (**Figure 3D**). Similarly, after applying the confidence threshold of 0.1, the interspecies disease domain structure collapses as only two links survive this constraint (**Figure 3E**). The level of confidence in relationships within the interspecies disease domain is estimated to be between 0.25 and 0.75 for both links and *RPN9* is the only gene found differentially expressed in the *Drosophila* dataset. This indicates that averaging the expression patterns for different gene transcripts reduces the information content of the network considerably and should be avoided for accurate prediction of the disease phenotype and generating biologically relevant regulatory networks.

Exhaustive Construction of Disease Network

We used the exhaustive Dandelion algorithm to overcome these limitations and provide a detailed interaction map of molecular pathology that extends our knowledge of disease mechanism across species. In contrast to the naïve variant, the exhaustive Dandelion algorithm searches the space of possible relationships at the level of gene transcripts to find the best scoring interspecies regulatory network. It can accommodate missing data and possible dissimilarities by identifying the best fit for a given relationship across species.

Bayesian networks which are generated using the exhaustive Dandelion algorithm can accurately predict the disease status from the expression levels of genes coding for proteasomal components (**Figure 4A**). We observe over 91% sensitivity and 80% specificity in the prediction of the disease phenotype in the human dataset (with an average SSE under 0.18), and similar values were obtained for the *Drosophila* and mouse datasets. The interspecies disease networks have very high predictive value for other species while they tend to avoid overfitting to a given dataset. This is evident from the low level of variation in SSE between constructed interspecies networks (0.06 in human, 0.11 in mouse, and 0.08 in *Drosophila*). The predictive ability of the interspecies models is highly robust towards the use of different organisms for training and testing, as the average SSE for a given species only slightly varies between different networks. Furthermore, the generated interspecies disease networks exhibit high sensitivity and specificity scores towards their informativeness to the prediction of the disease status. The majority of these networks provide sensitivity and specificity scores higher than 70% (**Figure 4B**). All constructed networks perform

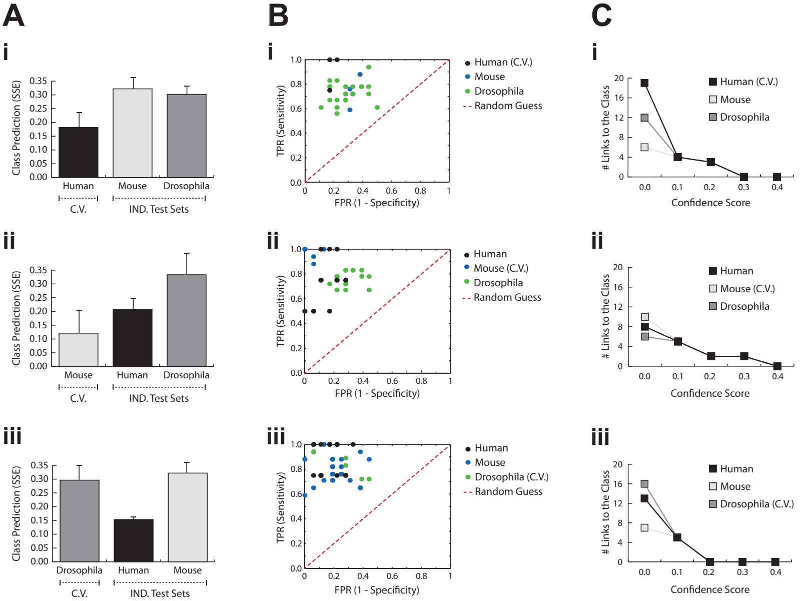


Figure 4 – Performance of the exhaustive Dandelion algorithm. A) The average Sum of Squared Error (SSE) for prediction of the disease phenotype (OPMD vs. control) given the gene expression profiles within the disease networks learnt on human (**i**), mouse (**ii**), or *Drosophila* (**iii**). The cross-validation set which is used during the training phase is depicted by *C.V.* and the independent test sets are grouped as *IND. Test Sets*. **B)** ROC space demonstrates the relative sensitivity and specificity of the generated networks in predicting the disease phenotype. The results from random expectations are illustrated by the red dash-line. **C)** Number of relationships between genes and the class node, after applying confidence thresholds, are depicted in line per species.

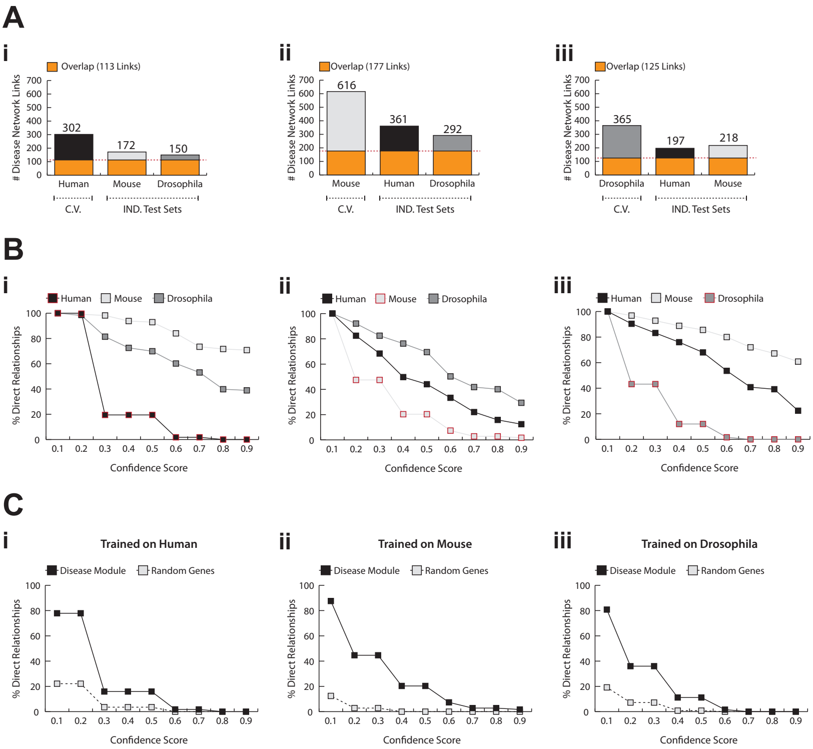


Figure 5 – Translatability and robustness of interspecies disease networks. A) The number of links that were found during interspecies translation and optimization of the disease networks per individual datasets. The red dash-line depicts the number and fraction of links that can be found in all species with the confidence threshold of 0.1. The translatability of disease networks learnt and trained on human (**i**), mouse (**ii**), and *Drosophila* (**iii**) are presented separately. The cross-validation set which is used during the training phase is depicted by *C.V.* and the independent test sets are grouped as *IND. Test Sets*. **B)** The translatability of relationships over series of different confidence thresholds. These line plots demonstrate the percentage of relationships with confidence score higher than the threshold. For the independent testing datasets the ratio is towards the number of links that were expected to be found after generation of the network map. **C)** The robustness of disease networks are assessed according to the level of connectivity for genes encoding for the proteasome as compared to the set of randomly selected genes at different confidence thresholds.

---

significantly better than random expectations, as presented in the ROC spaces (**Figure 4B**). In addition, the gene networks are strongly connected to the class node (representing information on the control and disease states of the samples) since the number of genes connected to the class node only drops to 0 when the confidence threshold was raised to 0.3, 0.4, or 0.2 for networks learnt on human, mouse, or *Drosophila*, respectively (**Figure 4C**). These are very restrained confidence thresholds as they require networks to share the same level of confidence for interactions across all species, and compare favorably to the low number of links remaining at the lower threshold of 0.1 with the naïve Dandelion algorithm.

Figure 5 demonstrates the level of robustness and translatability of the obtained disease networks. A large fraction of relationships (37.4% in human, 28.7% in mouse, and 34.3% in *Drosophila*) can be translated and found in the interspecies disease network with the confidence threshold of 0.1 (**Figure 5A**). Remarkably, an average of more than 60% of the translated links can be found in all

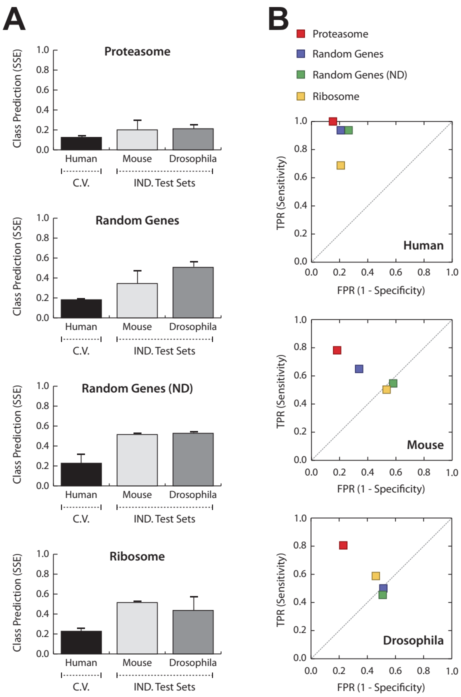


Figure 6 – Specificity of the proteasome towards prediction of disease states. A) The average Sum of Squared Error (SSE) for prediction of the disease phenotype (OPMD vs. control) given the gene expression profiles within the constructed networks learnt on the proteasome, 100 random genes, 70 not-deregulated random genes (ND), and the ribosome. The cross-validation set which is used during the training phase is depicted by *C.V.* and the independent test sets are grouped as *IND. Test Sets*. **B)** ROC space demonstrates the relative sensitivity and specificity of the generated networks in predicting the disease phenotype. The proteasome, 100 random genes, 70 random genes (ND), and ribosome are illustrated in different colors (red, purple, green, and yellow, respectively). The results from random expectations are illustrated by the gray dash-line.

organisms. It is evident that the intraspecies networks are highly resistant towards noise and the range of confidence in which interactions can be found in the training set is at least 0.7 and are as high as 0.9 in *Drosophila* and mouse datasets (**Figure 5B**). This value is even higher for relationships that are successfully translated from the intraspecies network to the other organisms (**Figure 5B**). Noticeably, the interspecies networks can still be obtained when applying a very stringent confidence threshold of 0.9 for all three constructed interspecies disease networks. More than 71% and 39% of translated relationships from human pass the confidence threshold of 0.9 in mouse and *Drosophila* datasets, respectively. However, a slightly more severe drop in translatability rate is observed for networks learnt on the mouse data. This can be expected due to the presence of overexpression and possibly other artifacts in this model system, also reflected by the higher level of interconnectivity of these networks. Despite the presence of noise and other artifacts in these datasets, a large fraction of interactions between genes encoding for the proteasome have high confidence scores in the interspecies networks (**Figure 5B**). This is not true for links associated with the randomly selected genes as the majority of those relationships do not pass the confidence threshold of 0.1 (**Figure 5C**). Overall, these results show model-driven selective and predictive ability of the exhaustive Dandelion algorithm in capturing the disease-related relationships between genes in which exhaustive Dandelion significantly outperforms the naïve Dandelion algorithm.

To assess the specificity of the proteasome in providing accurate prediction of the disease status, we compared the SSE, sensitivity, and specificity of the networks learnt on the proteasome to that of three additional gene sets. The exhaustive Dandelion algorithm was applied to a set of 70 random genes from which none is deregulated (ND) in OPMD, a set of 100 randomly selected genes containing also deregulated genes that are expected to link with the class node in one species but not necessarily across species, and 87 genes coding for the structurally-related ribosomal proteins, which are not known to be consistently differentially expressed in different species (Anvar et al., 2011). Noticeably, interspecies networks constructed on the proteasome significantly outperformed (86% sensitivity and 81% specificity across species) those constructed on other gene sets (**Figure 6**). Strikingly, the predictive accuracy of networks learnt on the proteasome was slightly improved from the previous experiment (**Figure 4**) in which additional 30 random genes

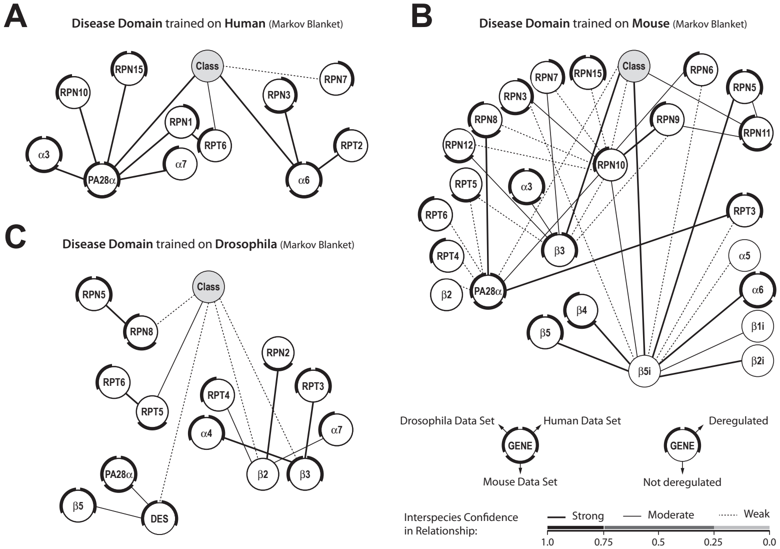


Figure 7 – Interspecies disease domains. These interspecies class network structures are learnt on human (**A**), mouse (**B**), or *Drosophila* (**C**) dataset and optimized across species. Class network structures are presented according to Markov blanket criteria. Nodes represent genes. The outer ring reflects deregulation in the expression in the different species (**a**, **b**). Relationships are depicted with lines that represent different degree of confidence in relationships (described in **c**).

---

were included. In contrast, the class prediction performance of the other networks was much lower. The class prediction error for networks learnt on the random genes was much higher than that of the proteasomal genes (average SSE of 0.43 and 0.21, respectively) but slightly lower than that of non-deregulated random genes and the ribosome (0.52, and 0.48, respectively) (**Figure 6A**). Although the performance is still acceptable for training and testing on human, the decrease in the level of sensitivity and specificity of non-proteasomal networks is particularly apparent during the translation phase (in this case from human data to mouse and *Drosophila*) (**Figure 6B**), indicating that the links between non-proteasomal genes are not conserved across the different species. Altogether, these results indicate a model-driven selective ability of the algorithm in capturing the most informative and consistent gene relationships which led to the construction of a highly robust interspecies disease network.

Network Genes and Identification of Key Regulators

Interspecies disease domains represent the most robust, disease-associated gene networks. They are identified by the class node (describing the disease status) and the associated Markov blanket of interactions with the confidence threshold of 0.1 across species (**Figure 7**). In the original experiment, the interspecies disease domain that is trained on human data shows the most robust network as the overall confidence in relationships is very high (**Figure 7A**). The mouse data, however, produced the highest number of relatively weaker relationships among genes (**Figure 7B**). The interspecies disease domain that is trained on the *Drosophila* data shows the same level of robustness as those constructed and trained on human (**Figure 7C**). In *Drosophila*, *Desmin* (*DES*), a randomly selected gene, is connected to the class node as part of the disease domain. Although *DES* (a muscle-specific class III intermediate filament) is a member of the random set, it is significantly deregulated in both human and *Drosophila* datasets. This gene has been clearly

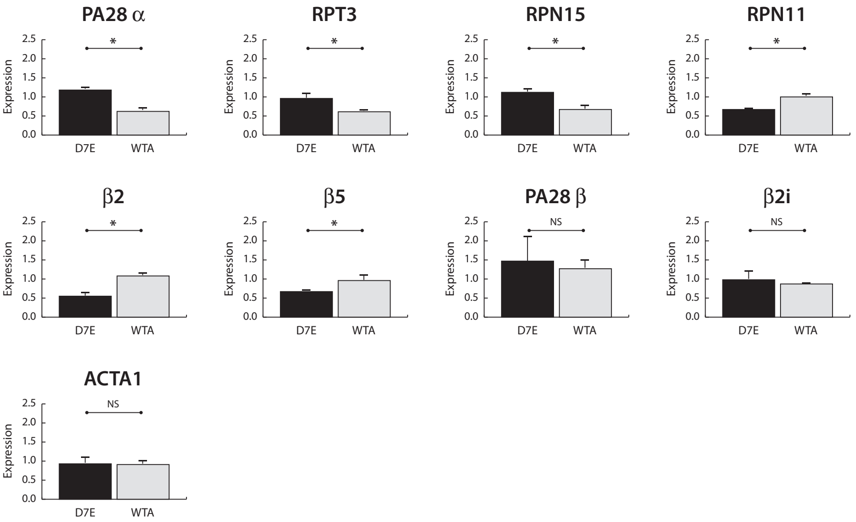


Figure 8 –Validation of differential expression of disease associated genes in an unseen disease model. Results from qPCR experiments measuring differences in gene expression between control cells (WTA, N=3 independent cultures) and cells expressing the OPMD-associated PABPN1 with expanded repeat (D7E, N=3 independent cultures). Expression levels were normalized to *Desmin* to correct for differences in the myogenicity in the different cell cultures. Significant differences ($P < 0.05$, Student´s T-test) between measured expression values in D7E and WTA cells are indicated by *, whilst NS stands for no significant difference. *PA28*α, *RPT3*, *RPN15*, *RPN11*, β*2*, and β*5* expression in IM2 cell lines were selected from the group of genes present in the interspecies disease domain. *PA28*β (deregulated in human dataset) was selected as its role in assembling the lid subunit of the immunoproteasome is highly similar to *PA28*α but not part of the interspecies disease domain. β*2i* is one of the two genes that remained connected to the class node in the interspecies disease domain constructed by naïve Dandelion approach. ACTA1 is a control for myotube formation.

linked to muscle differentiation (Capetanaki et al., 1997) and is likely associated with the OPMD phenotype. No other randomly selected genes appear in the disease network which indicates the reliability and the specificity of the obtained networks. Overall, the interspecies disease domains exhibit a high level of robustness and informativeness towards different states of the disease. This is due to the presence of relationships that can be translated across species with at least a moderate confidence (91.7% in human, 55.3% in mouse, and 71.4% in *Drosophila*). Moreover, the interspecies disease domains contain a large number of nodes that are differentially expressed in at least one species (100% in human, 80% in mouse, and 92.9% in *Drosophila*). Furthermore, the majority of genes are shared between at least two interspecies disease domains (81.8%, 64%, and 78.6%, for disease domains after training on human, mouse and *Drosophila,* respectively). Many of the links between genes present in these network structures demonstrate a strong correlation in expression profiles in the different species (**Table S4** in **Text S1**). Overall, these results indicate that the expression levels of the majority of genes in the constructed interspecies networks are strongly correlated and more likely to be associated with the OPMD phenotype than genes that are differentially expressed in single species.

Evaluation of Disease Networks on Unseen Disease Model

The model-driven and interspecies selection of genes that are most likely to be associated with the disease phenotype suggests their association with the disease in an independent and unseen disease model. Therefore, we evaluated the disease-related transcriptional changes for a subset of genes (selected from the interspecies disease domains) in the IM2 cell model (Raz et al., 2011) with moderate overexpression of the wild-type PABPN1 (WTA) or the mutant PABPN1 protein isoform (D7E). Remarkably, all the selected genes (*PA28α*, *RPT3*, *RPN15*, *RPN11*, *β2*, and *β5*) showed significant differential expression in an unseen IM2 cell model (**Figure 8**). *PA28α* appears to be an essential hub in the interspecies disease domains trained on the human and mouse datasets (**Figure 7**). Noticeably, it is also significantly deregulated between D7E and WTA (**Figure 8**). In contrast, *PA28β*, which is a closely related homolog in the PA28 complex (Rechsteiner and Hill, 2005) and also significantly deregulated in human dataset, do not play a part in the interspecies disease domains. Interestingly, it is evident that the expression pattern of *PA28β* is not deregulated between the D7E and WTA cells (**Figure 8**). Next, we assessed the expression of the *β2i*, a member of immunoproteasome core subunit, present in the interspecies disease domain constructed with the naïve Dandelion algorithm. This gene is not differentially expressed between D7E and WTA cells (**Figure 8**). Overall, these results highlight the unique ability of the exhaustive Dandelion algorithm to identify disease-related genes that can be found across different OPMD model systems and patients.

DISCUSSION

Integration of transcriptome data from different species is far from trivial and is complicated by our limited knowledge of true protein orthologues and transcript variants coding for proteins with similar functions. Moreover, the presence of noise and artifacts specific to certain model systems usually leads to limited overlap between results obtained in cross-species comparisons (Lu et al., 2009; Zhou and Gibson, 2004; Oliva et al., 2005; Blake et al., 2003). In this paper, we developed a Bayesian-based methodology (Dandelion algorithm) to model gene networks associated with the same disease in different species. We showed that the integration and analysis of gene expression datasets from various species increase the robustness of the constructed networks and the predictive accuracy of the disease state. We also demonstrated that the interspecies translation of the networks helps to avoid overfitting. A newly developed model-driven selection of transcripts that are most likely to be coding for orthologous proteins is essential for the generation of robust interspecies disease networks.

Our approach for Bayesian modeling of datasets on a similar phenotype from different model systems and patients is rather unique. Several approaches have been described to avoid overfitting and increase the robustness of Bayesian networks. For example, informative priors derived from protein-protein interaction (PPI) data or from the literature have been used to generate more stable and biologically meaningful networks (Segal et al., 2003; Pe'er et al., 2002; Steele et al., 2009; Jansen et al., 2003). While these methods obviously bias the results towards well-known regulatory interactions (Sprinzak et al., 2003; Joyce and Palsson, 2006), these methods may ultimately be combined with our modeling approach to obtain regulatory networks with a more straightforward biological interpretation.

Our method was applied to an *a priori* defined gene module coding for a well-known biological structure, the proteasome. Several studies in *S. cerevisiae* (Zhang et al., 2005; Tanay et al., 2004; Luscombe et al., 2004; Han et al., 2004) have demonstrated the value of an integrative modeling approach providing modularized interaction networks without prior assumptions. Zhang et al. (2005), for instance, took an approach in which they integrated a number of different available data sources, from PPIs to sequence homology and gene co-expression, while Tanay et al. (2004) and others (Luscombe et al., 2004; Han et al., 2004) expanded on the statistical analysis of network properties and identifying modules within the network structure. The performance of these models depends on the availability of high quantities of samples and may be prone to overfitting due to the presence of noise and other model-specific artifacts. Therefore, a combination with our interspecies translation approach may enable the allowing of larger gene regulatory networks with multiple gene modules and connections between them.

In this study, three microarray datasets from *Drosophila*, mouse and human, that are all concerned with OPMD, are used to gain insight into key regulatory relationships of interspecies disease networks that are directly and robustly associated with the disease. Previously, we have established the importance of the deregulation of the ubiquitin-proteasome system (UPS) for the disease etiology (Anvar et al., 2011). From the different components of the UPS, the down-regulation of the proteasome has been associated with the late-onset of the disease (Anvar et al., 2011) as the reduced proteasome activity can lead to futile protein degradation. However, little is known about the key components of the proteasome that are contributing to the OPMD phenotype. Hence, the generation of interspecies disease networks for the proteasome encoding genes now shed some light on the underlying regulatory mechanisms that govern the disease-related transcriptional changes of the proteasome encoding genes.

We identified PA28α, one of the three components of the PA28 subunit, as an important hub gene in the interspecies disease domain and validated its significant differential expression in an unseen disease model. PA28α plays an important role in assembling the lid subunit of the immunoproteasome and stimulating the proteasome core component (Rechsteiner and Hill, 2005). Previously we showed that the induction of immunoproteasome activity leads to a significant reduction in the nuclear expPABPN1 accumulation (Anvar et al., 2011). This observation further signifies the role of PA28 assembly and the immunoproteasome in the disease etiology. In contrast, the other PA28 component PA28β although significantly deregulated in human OPMD patients, appears to play a less crucial role since its association with the disease did not translate to the OPMD animal models and could not be reproduced in the OPMD cell model system. On the other hand, the association of β2 and β5, members of the proteasome core subunit, with the disease was identified by the interspecies disease domains and reproduced in the OPMD cell model. Down-regulation of the proteasome core subunit can lead to futile protein degradation which results in protein accumulation. Our analysis suggests that β2 and β5 are vital regulators of the proteasome activity which are disease associated. It has been shown that the down-regulation of the proteasome core subunit can trigger expPABPN1 accumulation and play a role in the disease late-onset (Anvar et al., 2011). Relevant to the late-onset of the OPMD, previously it has been shown that the proteasome activity declines during muscle ageing (Ferrington et al., 2005; Combaret et al., 2009; Lee et al., 1999), a phenomena which is highly associated with the transcriptional changes of the proteasomal genes (Lee et al., 1999). In follow-up studies, the functional role of proteasomal protein dysregulation in the disease pathology and ageing of muscles needs to be investigated. Furthermore, the functional relevance of gene regulatory relationships should be investigated where changes in protein level mimic the *in vivo* situation and directly affect the protein catabolism. This would ultimately result in better understanding of the mechanism in which the loss of proteostasis leads to degenerative loss of muscle function during ageing and in OPMD.

In conclusion, this study presents a state-of-the-art strategy in constructing interspecies disease networks that provide crucial and comprehensive information on gene regulatory relationships. This leads to better understanding and identification of the molecular mechanisms underlying the disease. The high level of specificity and sensitivity of these models enables the prioritization of candidate regulators of molecular disease mechanisms to be studied in follow-up validation experiments. In particular, it is crucial to carry out additional experiments to investigate the functional relevance of proteasomal proteins dysregulation to the OPMD pathology. We believe that robust and unbiased construction of the interspecies networks for rare or complex human diseases can lead to novel discovery and identification of key regulators which can ultimately offer potential targets for therapeutic interventions and drug developments.

Acknowledgement

We are grateful to Drs. Jelle Goeman (Leiden University Medical Center, The Netherlands) and Martijn Schuemie (Erasmus Medical Center and Leiden University Medical Center, The Netherlands) for their expertise, assistance and valuable discussions on designing the experiments.

Reference List

Anvar,S.Y., 't Hoen,P.A., and Tucker,A. (2010). The identification of informative genes from multiple datasets with increasing complexity. BMC. Bioinformatics. *11*, 32.

Anvar,S.Y., 't Hoen,P., Venema,A., van der Sluijs,B., van Engelen,B., Snoeck,M., Vissing,J., Trollet,C., Dickson,G., Chartier,A., Simonelig,M., van Ommen,G.J., van der Maarel,S., and Raz,V. (2011). Deregulation of the ubiquitin-proteasome system is the predominant molecular pathology in OPMD animal models and patients. Skeletal Muscle *1*, 15.

Barabasi,A.L., Gulbahce,N., and Loscalzo,J. (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet. *12*, 56-68.

Blake,W.J., KAErn,M., Cantor,C.R., and Collins,J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633-637.

Brais,B., Bouchard,J.P., Xie,Y.G., Rochefort,D.L., Chretien,N., Tome,F.M., Lafreniere,R.G., Rommens,J.M., Uyama,E., Nohira,O., Blumen,S., Korczyn,A.D., Heutink,P., Mathieu,J., Duranceau,A., Codere,F., Fardeau,M., and Rouleau,G.A. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet *18*, 164-167.

Capetanaki,Y., Milner,D.J., and Weitzer,G. (1997). Desmin in muscle formation and maintenance: knockouts and consequences. Cell Struct. Funct. *22*, 103-116.

Chartier,A., Benoit,B., and Simonelig,M. (2006). A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. EMBO J *25*, 2253-2262.

Chartier,A., Raz,V., Sterrenburg,E., Verrips,C.T., van der Maarel,S.M., and Simonelig,M. (2009). Prevention of oculopharyngeal muscular dystrophy by muscular expression of Llama single-chain intrabodies in vivo. Hum. Mol. Genet. *18*, 1849-1859.

Combaret,L., Dardevet,D., Bechet,D., Taillandier,D., Mosoni,L., and Attaix,D. (2009). Skeletal muscle proteolysis in aging. Curr. Opin. Clin. Nutr. Metab Care *12*, 37-41.

Davies,J.E., Wang,L., Garcia-Oroz,L., Cook,L.J., Vacher,C., O'Donovan,D.G., and Rubinsztein,D.C. (2005). Doxycycline attenuates and delays toxicity of the oculopharyngeal muscular dystrophy mutation in transgenic mice. Nat Med. *11*, 672-677.

Fan,X. and Rouleau,G.A. (2003). Progress in understanding the pathogenesis of oculopharyngeal muscular dystrophy. Can. J. Neurol. Sci. *30*, 8-14.

Ferrington,D.A., Husom,A.D., and Thompson,L.V. (2005). Altered proteasome structure, function, and oxidation in aged muscle. FASEB J. *19*, 644-646.

Fielding,A.H. (2007). Introduction to classification. In Cluster and classification techniques for the Biosciences, Cambridge University Press), p. 86.

Friedman,N. (2004). Inferring cellular networks using probabilistic graphical models. Science *303*, 799-805.

Friedman,N., Geiger,D., and Goldszmidt,M. (1997). Bayesian network classifiers. Machine Learning *29*, 131-163.

Friedman,N., Linial,M., Nachman,I., and Pe'er,D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601-620.

Goeman,J.J., van de Geer,S.A., de,K.F., and van Houwelingen,H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. *20*, 93-99.

Goldstein,D.B. (2009). Common genetic variation and human traits. N. Engl. J. Med. *360*, 1696-1698.

Han,J.D., Bertin,N., Hao,T., Goldberg,D.S., Berriz,G.F., Zhang,L.V., Dupuy,D., Walhout,A.J., Cusick,M.E., Roth,F.P., and Vidal,M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature *430*, 88-93.

Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F., and Gerstein,M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449-453.

Joyce,A.R. and Palsson,B.O. (2006). The model organism as a system: integrating 'omics' data sets. Nat Rev Mol. Cell Biol. *7*, 198-210.

Karlebach,G. and Shamir,R. (2008). Modelling and analysis of gene regulatory networks. Nat Rev Mol. Cell Biol. *9*, 770-780.

Kluger,Y., Yu,H., Qian,J., and Gerstein,M. (2003). Relationship between gene co-expression and probe localization on microarray slides. BMC. Genomics *4*, 49.

Lee,C.K., Klopp,R.G., Weindruch,R., and Prolla,T.A. (1999). Gene expression profile of aging and its retardation by caloric restriction. Science *285*, 1390-1393.

Lu,Y., Huggins,P., and Bar-Joseph,Z. (2009). Cross species analysis of microarray expression data. Bioinformatics. *25*, 1476-1483.

Luscombe,N.M., Babu,M.M., Yu,H., Snyder,M., Teichmann,S.A., and Gerstein,M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308-312.

Murphy,K.P. (2001). The Bayes Net toolbox for Matlab. Computing Science and Statistics: Proceedings of the Interface *33*, 331-350.

Oliva,A., Rosebrock,A., Ferrezuelo,F., Pyne,S., Chen,H., Skiena,S., Futcher,B., and Leatherwood,J. (2005). The cell cycle-regulated genes of Schizosaccharomyces pombe. PLoS. Biol. *3*, e225.

Pache,R.A., Zanzoni,A., Naval,J., Mas,J.M., and Aloy,P. (2008). Towards a molecular characterisation of pathological pathways. FEBS Lett. *582*, 1259-1265.

Pe'er,D., Regev,A., and Tanay,A. (2002). Minreg: inferring an active regulator set. Bioinformatics. *18 Suppl 1*, S258-S267.

Pearl,J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. (San Francisco: Morgan Kaufmann).

Pedraza,J.M. and van Oudenaarden,A. (2005). Noise propagation in gene networks. Science *307*, 1965-1969.

Raj,A. and van Oudenaarden,A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216-226.

Raz,V., Routledge,S., Venema,A., Buijze,H., van der Wal,E., Anvar,S.Y., Straasheijm,K.R., Klooster,R., Antoniou,M., and van der Maarel,S.M. (2011). Modeling Oculopharyngeal Muscular Dystrophy in Myotube Cultures Reveals Reduced Accumulation of Soluble Mutant PABPN1 Protein. Am. J. Pathol.

Rechsteiner,M. and Hill,C.P. (2005). Mobilizing the proteolytic machine: cell biological roles of proteasome activators and inhibitors. Trends Cell Biol. *15*, 27-33.

Schadt,E.E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature *461*, 218-223.

Schwarz,G. (1978). Estimating the dimension of a model. The Annals of Statistics *6*, 461-464.

Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D., and Friedman,N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. *34*, 166-176.

Shahrezaei,V. and Swain,P.S. (2008). The stochastic nature of biochemical networks. Curr. Opin. Biotechnol. *19*, 369-374.

Sprinzak,E., Sattath,S., and Margalit,H. (2003). How reliable are experimental protein-protein interaction data? J. Mol. Biol. *327*, 919-923.

Steele,E., Tucker,A., 't Hoen,P.A., and Schuemie,M.J. (2009). Literature-based priors for gene regulatory networks. Bioinformatics. *25*, 1768-1774.

Stone,M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society *36*, 111-147.

Tanay,A., Sharan,R., Kupiec,M., and Shamir,R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc. Natl. Acad. Sci. U. S. A *101*, 2981-2986.

Trollet,C., Anvar,S.Y., Venema,A., Hargreaves,I.P., Foster,K., Vignaud,A., Ferry,A., Negroni,E., Hourde,C., Baraibar,M.A., 't Hoen,P.A., Davies,J.E., Rubinsztein,D.C., Heales,S.J., Mouly,V., van der Maarel,S.M., Butler-Browne,G., Raz,V., and Dickson,G. (2010). Molecular and phenotypic characterization of a mouse model of oculopharyngeal muscular dystrophy reveals severe muscular atrophy restricted to fast glycolytic fibres. Hum. Mol. Genet. *19*, 2191-2207.

Zhang,L.V., King,O.D., Wong,S.L., Goldberg,D.S., Tong,A.H., Lesage,G., Andrews,B., Bussey,H., Boone,C., and Roth,F.P. (2005). Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J. Biol. *4*, 6.

Zhou,X.J. and Gibson,G. (2004). Cross-species comparison of genome-wide expression patterns. Genome Biol. *5*, 232.

APPENDIX

Table S1 – Terminological definitions.

Term	Definition
Disease Module	Molecular pathway in which gene expression profiles are significant associated with the disease phenotype. Modules are described based on the current KEGG (Kyoto Encyclopedia of Genes and Genomes) annotation of molecular pathways.
Intraspecies Network	Gene network in which structural relationships among genes are based on the training with data from a single organism.
Interspecies Network	Gene regulatory network of which the structure holds a consensus across all species.
Sum Squared Error	The SSE measurement is the sum of the squares of the deviations between the measured expression values (or assigned disease phenotype) and the values predicted from the response variable which can be the class node (discrete variable), gene or gene transcript node (continuous variable). The identifier for the graph node is represented by *g* and the case id is represented by *i*. $SSE_g = \sum_{i=1}^{n} \left(measured\ value_{g,i} - predicted\ value_{g,i}\right)^2$
Sensitivity	The probability of accurate prediction of cases with the disease-associated phenotype. $Sensitivity = \frac{number\ of\ True\ cases^{\ disease}}{total\ number\ of\ True\ cases^{\ disease}}$
Specificity	The probability of accurate prediction of control cases without the disease-associated phenotype. $Specificity = \frac{number\ of\ True\ cases^{\ control}}{total\ number\ of\ True\ cases^{\ control}}$
Confidence Score	The ratio of the number of times a link is found in a network structure to the maximum number of times the link can be found. For the training set (species *A*): $Confidence\ Score = \frac{number\ of\ times\ a\ link\ is\ found\ \left(n_{species\ A}\right)}{total\ number\ of\ constructed\ networks\ on\ A}$ For the independent test set (species *B*): $Confidence\ Score = \frac{number\ of\ times\ a\ link\ is\ found\ \left(n_{species\ B}\right)}{n_{species\ A} \times total\ number\ of\ constructed\ networks\ on\ B}$
Robustness	The number of relationships found for genes from the disease module compared to those from random genes after applying different confidence thresholds.

Translatability	The likelihood of finding genes neighboring relatives that are selected as part of the intraspecies network structure during the phase of independent testing in the other species.
Naïve Dandelion	A class of Dandelion algorithm in which the networks are constructed on datasets derived from different organisms, where transcript expression levels for the same gene are averaged.
Exhaustive Dandelion	A class of Dandelion algorithm in which the structure of intraspecies networks are learnt on gene transcript level. This procedure involves a model-driven selection of the most probable homologous transcript isoform which is best translated across species.
Disease Domain	A sub-network structure associated with the class (disease) node which is defined based on the Markov blanket principle for the extension of the class node connectivity. This sub structure is composed of class node, its children, and its children's other parents that share the same level of confidence (≥0.1). A Markov blanket of the class node is the only knowledge needed to predict the disease phenotype.

Protocol S1 – Algorithm for Simulated Annealing Structure Learning.

```
Input:   $t_0 = 10, maxfc = 1000, D, mode, netmap$
         $fc = 0, t = t_0, t_n = 0.001$
         $c = (t_n / t_0)^{1/maxfc}$
         Initial bn to a Bayesian classifier with no inter-gene links
         $result = bn$
         $oldscore = score(bn)$
         While $fc < maxfc$ do
            For each operator do
                  If $mode =$ 'train'
                           Apply operator to bn
                  Else if $mode =$ 'test'
                           Apply operator to bn based on links available in networkMap
                  End if
                  $newscore = score(bn)$
                  $fc = fc + 1$
                  $dscore = newscore - oldscore$
                  If $newscore > oldscore$ then
                           $result = bn$
                  Else if $r(0,1) < e^{dscore/t}$ then
                           Undo the operator
                  End if
            End for
            $t = t X c$
         End while
Output:  result
```

Protocol S2 – Dandelion algorithm of interspecies construction of disease network.

```
Input:   $Species_{train}$, $\{Species_{test\ 1}, \ldots, Species_{test\ M}\}$, $train_{folds}$, $\{test_{folds\ 1}, \ldots, test_{folds\ M}\}$, $exhaustive_{T/F}$
         For $k = 1$ to $train_{folds}$
             Learn $intraspeciesTranscript_{bn}$ using Algorithm 1 on training folds of $Species_{train}$
             Score $Scpecies_{train}\{Nodes_{SSE}, Nodes_{STD}, Links_{Confidence}\}$
             If $exhaustive = true$
                 Transform $intraspeciesTranscript_{bn}$ to $intrascpeciesGene_{bn}$
             End if
             Assess Disease Connection
             If $intraspeciesGene_{bn}$ is not connected to $disease\ node$ then
                 Drop $intrascpeciesGene_{bn}$
             Else
                 Translate $intrascpeciesGene_{bn}$ to $networkMap$
                 For $i = 1$ to $M$
                     Optimize and Test $networkMap$ in $Species_{test\ i}$ using Algorithm 1
                     Score $Scpecies_{test\ i}\{Nodes_{SSE}, Nodes_{STD}, Links_{Confidence}\}$
                 End for
             End if
         End for
         Integrate $intraspeciesGene_{bn}$ using $Links_{Confidence}$ threshold of $0.1$
Output:  $interspecies_{bn}$
```

Table S2 - Gene lists for independent tests and performance assessments.

Proteasome and 30 Random Genes		100 Random Genes		70 Random Genes (not deregulated)		Ribosome	
PSMD3	LOC643791	LOC644993	LOC651979	CPSF4L	WTAP	FAU	RPS6
PSMD12	C9orf79	LOC147710	OR4A47	LOC652683	CRTC2	RPSA	RPS7
PSMD11	MGRN1	PCDHB5	KCTD14	MME	LSM14B	RPL10A	RPS9
PSMD6	LOC653587	KIAA1688	CDK5RAP2	LOC653261	PRKG2	RPL3	RPS10
PSMD7	CNGA4	A4GALT	TMPRSS4	CD200R1	LUM	RPL3L	RPS11
PSMD13	OTOR	SFN	ADAMTS13	HSD11B1	PRUNE	RPL4	RPS12
PSMD14	GPR89A	BCL10	FRAS1	PDE4DIP	RPS3AP47	RPL5	RPS13
PSMD8	GPR89B	MSX2	SCUBE1	EEPD1	P2RX2	RPL6	RPS14
SHFM1	HAPLN4	SNRPB	LOC642855	KRTAP4-11	NAV1	RPL7	RPS15
PSMD4	LOC641994	HERC3	LOC442261	SLFN14	XRCC2	RPL7A	RPS15A
PSMD2	THBS2	HRASLS2	ZNF100	POU4F1	C17orf87	RPL8	RPS16
PSMD1	ZNF768	DLD	HDGFRP3	LOC442132	CACNA1I	RPL9	RPS17
PSMC2	KIAA1147	LOC649217	LOC642453	ST6GLA2	ELSPBP1	RPL11	RPS18
PSMC1	C19orf59	IGHG1	RHBDD1	ACTR3B	EPGN	RPL12	RPS19
LOC643668	BARHL2	GNPTAB	RSL1D1	PEF1	LOC650933	RPL13	RPS20

PSMC5
PSMC6
PSMC3
PSMC4
PSMA6
PSMA2
PSMA4
PSMA8
PSMA7
PSMA5
PSMA1
PSMA3
PSMB6
PSMB7
PSMB3
PSMB2
PSMB5
PSMB1
PSMB4
PSME1
PSME2
PSME3
PSME4
POMP
PSMF1
IFNG
PSMB9
PSMB10
PSMB8
PSMB11
AKR1CL1
CHRNA5
UNC13B
DES
STT3A

LOC400831
HMGN4
TSSK4
RTKN2
RXRA
MYL5
UBTD1
OR1J4
TRAPPC5
ADAM20

NOC4L
PLD3
LOC648974
GTPBP8
LIF
LOC440104
WAC
KALRN
UNC93A
IFNAR1
NMT1
LOC652750
LOC653707
SLC26A9
ETFDH
ADAM23
FBXO9
LOC643089
ATP5D
CST6
RPL11
FAM47B
LHFPL4
MGC42105
STOX2
FRMD5
CHL1
UNQ830
STCH
B4GALNT3
SUMO2
C20orf30
CNIH3
DBX2
GSTM5P1

LOC652610
LOC646699
KNDC1
DACT3
FLJ16369
VIPR1
COPS8
NIF3L1
PPAP2C
LOC644431
TCTE3
TTF2
RPS7
ITGA8
CCAR1
PDCD10
LOC651400
CDC42BPG
SP2
LOC649432
LOC732093
TMEM165
LHCGR
SPAG7
INOC1
OR2T10
DEPDC5
ADAD1
LOC339529
FZD9
CD46
JARID1B
DUX4
DPPA4
YSK4

OGG1
TAF9B
LOC653421
LOC441347
FRMPD2
HSCB
CHD1
LOC645781
LOC729446
FAM129C
FAM90A15
C1orf187
HIPK2
XKR3
RAB2A
FOXR1
CD72
TRAF4
NCAN
HRC
LOC643577
AKR7A2P1
PLK2
RABL2B
CLGN
LRRC49
CHORDC1
KRT18P51
OR13G1
CCL21
LRFN2
SLC35A5
RDH12
FAM154B
LOC388948

HDX
APOL3
CNOT4
PFAS
MAP3K14

RPL15
RPL17
RPL18
RPL18A
RPL19
RPL21
RPL22
RPL23A
RPL24
RPL26
RPL27
RPL30
RPL27A
RPL28
RPL29
RPL31
RPL32
RPL34
RPL35A
RPL36AL
RPL37
RPL37A
RPL38
RPL39
RPL41
RPL36A
RPLP0
RPLP1
RPLP2
RPS2
RPS3
RPS3A
RPS4X
RPS4Y1
RPS5

RPS21
RPS23
RPS24
RPS25
RPS26
RPS27
RPS27A
RPS28
RPS29
UBA52
RPL14
RPL23
RPL35
RPL13A
RPL36
MRPL13
RPS27L
RPL26L1
C15orf15
RPL10L
RPL22L1
RSL24D1P11

Table S3 – The list of primers that were used for qPCR validation study in IM2 cell model of OPMD.

Gene	FW Primer Sequence	RV Primer Sequence
RPN11 (Psmd14)	CACCTGAACAGCTGGCAATA	GAGCATTGGGAACGAAGAAG
RPN15 (Shfm1)	AGCACGGCTACAAGATGGAG	TGAACCAAAAAGATTAAATCAAAACA
RPT3 (Psmc4)	ACCTCAGACCAGAAGCCAGA	CACCACACGGATAAATGCAG
β2 (Psmb7)	GCACTACCGCTGTCCTCACCG	AGGGGTGGTATGCACCCCGAG
β5 (Psmb5)	CGGTCGCAGCAGCCTCCAAA	GCATACACGGAGCCAGAGCCC
PA28α (Psme1)	AAGCCAAGGTGGATGTGTTC	GGGTACTGGGATGTCCAATG
PA28β (Psme2)	CCTGGAGAGTGAAAGCGAAA	GTCATCAGCCTCCTGGAAAA
β2i (Psmb10)	ATTTGCTCCTGGAACCACAC	CCACTTCATTCCACCTCCAT
ACTA1 (Acta1)	CGAGGTATCCTGACCCTGAA	AGGTGTGGTGCCAGATCTTC
mHPRT	CGTCGTGATTAGCGATGATG	TTTTCCAAATCCTCGGCATA

Table S4 – Correlation between the expression profiles of genes selected from the interspecies disease domains.

			Interspecies Confidence	Human		Mouse		Drosophila	
Gene A	**Gene B**	**Train Set**		**Correlation Score**	**P-value**	**Correlation Score**	**P-value**	**Correlation Score**	**P-value**
PA28α	RPN1	Human	Strong	0.6059	2.80E-03	0.8650	8.50E-11	-0.3762	2.37E-02
PA28α	RPN15	Human	Strong	-0.0466	8.37E-01	0.7521	4.50E-07	0.4801	3.03E-03
PA28α	RPN8	Mouse	Strong	-0.0670	7.57E-01	0.9245	1.58E-14	0.4658	4.20E-03
PA28α	RPT3	Mouse	Strong	0.5984	3.26E-03	0.7944	3.42E-08	0.4491	6.00E-03
PA28α	α3	Human	Strong	-0.5234	1.24E-02	0.8988	1.24E-12	0.5498	5.14E-04
RPN10	RPN3	Mouse	Moderate	0.5450	8.72E-03	0.6400	6.07E-05	0.5503	5.06E-04
RPN10	RPN6	Mouse	Moderate	0.6300	1.68E-03	0.8634	1.00E-10	0.8180	1.12E-09
RPN10	RPN9	Mouse	Strong	0.3338	1.29E-01	0.8760	2.46E-11	0.8545	3.36E-11
β3	RPN12	Mouse	Moderate	0.4635	2.98E-02	0.8653	8.15E-11	0.5090	1.52E-03
β3	RPN7	Mouse	Moderate	-0.2445	2.73E-01	0.9386	7.17E-16	0.6361	3.05E-05
β3	RPT3	Drosophila	Strong	0.4274	4.72E-02	0.8652	8.32E-11	0.5855	1.76E-04
β3	α3	Mouse	Moderate	0.0910	6.87E-01	0.9456	1.15E-16	0.8367	2.07E-10
β3	α4	Drosophila	Strong	0.2359	2.91E-01	0.9378	8.65E-16	0.7959	6.56E-09
β5i	α6	Mouse	Strong	-0.5980	3.28E-03	0.6593	3.01E-05	0.6393	2.70E-05
β5i	β1i	Mouse	Moderate	0.6729	6.00E-04	0.9053	4.63E-13	-0.1952	2.54E-01
β5i	β4	Mouse	Strong	0.1983	3.76E-01	0.7679	1.83E-07	0.6416	2.48E-05

perspectives

The rapid development of "omics" technologies (genomics, transcriptomics, proteomics, metabolomics, and others) has allowed for a more detailed understanding of complex biological systems. However, analytical approaches for meaningful interpretation of multilayer omics datasets are lagging behind the technological advancements. The sparse, amorphous, distributed, and poorly reproducible state of omics datasets and the lack of standard for generating such data in many disciplines further complicate the analysis of these datasets (Ioannidis, 2005). These bottlenecks pose the importance of developing stringent computational strategies and a validation regime that can distinguish between true signals and noise (Ioannidis and Khoury, 2011). Interdisciplinary approaches are required to bridge the growing gap between technological development, biomedical research, and computational biology. Thus, converging loops between theory and experiment should help to understand the dynamics of biological networks and processes. Integrative analyses of different and multifaceted biological datasets should facilitate the study of human genetic disorders.

In this Outlook, I firstly discuss the scientific rationale for carrying out multi-disciplinary research on a rare human genetic disorder and further outline how that can benefit society. I then describe and provide an overview on some of the key mechanistic insights unravelled by my colleagues and myself through the course of our studies. Finally, some directions which could enhance the evolution of the field of systems biology are discussed.

Remarkable insights from a rare event

DNA sequence polymorphisms contribute to individual differences in disease susceptibility. As genetic information can be passed onto mRNA and proteins that perform cellular functions, genetic studies have often focused on the one-to-one relationship between phenotype and genomic variation as the basis for knowledge discovery. There are, however, common patterns that underlie the diversity, complexity, development and progression of genetic diseases. Hence, looking for shared molecular activities, across organisms or processes with similar characteristics (such as ageing, late-onset neuromuscular and protein aggregation disorders), can lead to uncovering new insights into mechanisms that are important for diverse biological conditions. In particular, such combined strategies would facilitate the study of rare diseases. There are an estimated 8,000 rare disorders, many of which are known to be of genetic origin (Stolk et al., 2006; Schieppati et al., 2008). Given the low prevalence of rare diseases, it is particularly difficult to employ traditional approaches and, therefore, they require special integrative efforts to improve discovery of underlying mechanisms. Notably, in spite of the low prevalence of each rare disease, about 30 million people are estimated to be affected by a rare disease in Europe (Kaplan and Laing, 2004). In the work presented in this thesis, my colleagues and I have mainly focused on developing computational approaches and conducting an interdisciplinary study to unravel novel associations based on shared functional features. Here, I discuss an outlook on how such strategies can lead to significant findings that benefit society. I do this based on the result of our investigations on Oculopharyngeal muscular dystrophy (OPMD).

PABPN1, the protein mutated in OPMD, regulates poly(A) tail length and RNA stability (Lemay et al., 2010; Kuhn et al., 2009). As such, PABPN1 plays an important role in a variety of cellular processes (Kuhn et al., 2009; Wahle, 1991; Wahle, 1995; Lemieux and Bachand, 2009; Calado et al., 2000; Apponi et al., 2010). It has been shown that manipulation of PABPN1 or expanded PABPN1 (expPABPN1) expression levels in muscle cells, i.e. high over-expression or complete gene knockdown, leads to muscle defects including muscle weakness and muscle atrophy, impaired cell growth and apoptosis, and impaired cell fusion (Apponi et al., 2010; Chartier et al., 2006; Davies et al., 2006; Trollet et al., 2010). We have shown that expPABPN1 expression in muscle fibres leads to substantial gene deregulation in OPMD patients and in OPMD model systems (Anvar et al., 2011a; Raz et al., 2011; Trollet et al., 2010). In the OPMD mouse model, by performing an integrative analysis empowered by a number of computational and data-mining methods, we reported muscle atrophy as the major contributor to muscle weakness. This was evident from both the reduction of muscle mass and loss of contractile force due to increased fibrosis, mitochondrial defects, oxidative stress, and deregulation of the ubiquitin-proteasome system (see Chapter **one**). In this mouse model, these observations were mainly restricted to the glycolytic fibres. Despite some pathological similarities between OPMD patients and the mouse model, muscle atrophy is rare among OPMD patients. It is possible that the severe atrophy in glycolytic fibres of OPMD mice is the result of the high overexpression of expPABPN1 rather than the mutation itself. To correct for potential artefacts, an integrative approach was designed in which microarray datasets from three different organisms were combined to gain insight in the common molecular pathways that underlie OPMD. The result of such an extensive strategy, presented in Chapter **two**, was the identification of the ubiquitin-proteasome system (UPS) as the most prominently deregulated molecular pathway in OPMD model systems and patients (Anvar et al., 2011a). Transcriptome studies in non-muscle cells expressing expPABPN1 did not reveal prominent deregulation of UPS genes (Corbeil-Girard et al., 2005). This suggests that the effect of expPABPN1 on UPS deregulation is specific to muscle cells or to post-mitotic cells. Deregulation of the UPS has also been reported in myotonic dystrophy type 1 (Vignaud et al., 2010) and in muscle atrophy in mice (Cao et al., 2005; Bodine et al., 2001; Sandri, 2008). In addition, altered UPS activity has been associated with muscle ageing (Combaret et al., 2009; Lee et al., 1999). Together, these studies suggest that muscle cell function is tightly regulated by the UPS.

The UPS is the main regulator of protein homeostasis (also referred to as proteostasis) and is involved in a wide spectrum of human diseases including cancer, neurodegenerative disorders and diabetes (Hoeller and Dikic, 2009; Liu et al., 2000; Combaret et al., 2009; Taillandier et al., 2004; Ciechanover and Brundin, 2003). To maintain protein homeostasis, it is essential to uphold balance between activities of protein quality-control machineries, the UPS and autophagy-lysosome (Powers et al., 2009). These machineries can adequately respond to damaged proteins and organelles through adjustment of the level of chaperones and proteases (Goldberg, 2003; Meusser et al., 2005; Guisbert et al., 2008; Morimoto, 2008; Ron and Walter, 2007). However, progressive exhaustion of these quality-control systems, owing to ageing (Hipkiss, 2006; Wang et al., 2009; Munch and Bertolotti, 2010; Ben-Zvi et al., 2009) or genetic mutation (Olzmann et al.,

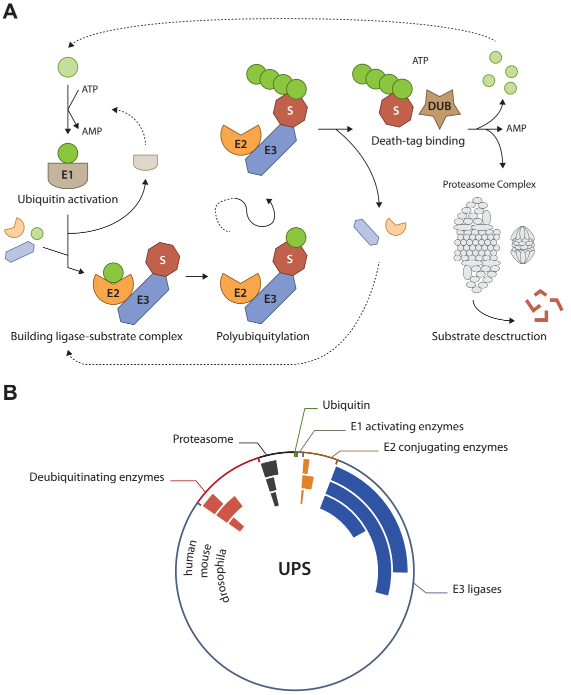


Figure 1 – Schematic overview of the ubiquitin-proteasome system. A) Protein degradation through the ubiquitin-proteasome system involves several steps. Firstly, the ubiquitin (Ub) is being activated by ubiquitin-activating enzyme (E1). Next, ubiquitin is delivered to ubiquitin-conjugation enzyme (E2) for formation of the E2-Ub, ubiquitin ligase (E3) and substrate complex. Consequently, ubiquitins are being transferred to the substrate in order to tag the substrate with the polyubiquitin chain. In the fourth step, E3 releases the polyubiquitylated substrate. The proteasome recognises the polyubiquitin chain as a degradation signal. Therefore, substrate is deubiquitinated and destroyed by the proteasome in ATP-manner. **B)** Within the ubiquitin-proteasome system, E2, E3, deubiquitinating enzymes and proteasome show significant deregulation. Pie charts illustrate the relative distribution of the deregulated genes widespread throughout different components of the ubiquitin-proteasome system across species. A fraction of deregulated genes within individual species are shown in dark colours.

2007), would lead to accumulation of altered proteins as the accumulation of misfolded proteins surpasses the system's capacity (Tyedmers et al., 2010). It has been suggested that the aggregation of proteins can spread to other proteins of mainly the same type (Gidalevitz et al., 2006; Rajan et al., 2001; Ben-Zvi and Goloubinoff, 2002) which could explain the age-dependent and progressive nature of protein aggregation phenomena. Excessive aggregation of proteins could lead to a progressive decline in the level of soluble proteins available in cells. This may result in reduced level of functional protein and, consequently, lead to pathophysiological abnormalities. Thus, understanding the molecular processes regulating cellular homeostasis may unravel mechanistic insights in pathological aspects of various protein aggregation and late-onset diseases.

The UPS involves an enzymatic cascade of ubiquitination and degradation steps. From the six UPS components, only E3-ligases, deubiquitinating enzymes and the proteasome were found to be consistently and prominently deregulated in OPMD model systems and patients (**Figure 1**). Particularly, E3-ligases are fundamental to the specificity of this system and are classified into the RING finger, HECT, and U-Box domains (Deshaies and Joazeiro, 2009). Moreover, many of the E3-ligases play an important role in maintaining genomic integrity (Lipkowitz and Weissman, 2011). Therefore, it is essential to understand what type of E3-ligases are involved in forming the E2-E3 complex to specify the fate of proteins involved in protein aggregation disorders. In OPMD, we found that a subset of the deregulated E3-ligases co-localize with the aggregates of mutant PABPN1. Moreover, their RNA expression profiles correlate with their sequential entrapment in intranuclear inclusion (INI) (Anvar et al., 2011a). It would be essential to look for possible E3-ligases that differentially bind and/or regulate wild-type and mutant PABPN1 in OPMD. This is important since differential regulation of PABPN1 may in part explain the enhanced level of PABPN1 aggregates and reduced level of soluble proteins in OPMD patients. Intriguingly, E3-ligases are also recognised as potential drug targets (Nalepa et al., 2006; Xu and Jaffrey, 2011).

Hence, focused profiling efforts can lead to the identification of ubiquitination events that are regulated by potential therapeutic compounds (Kim et al., 2011; Emanuele et al., 2011).

The process of ubiquitination and degradation through UPS machinery is a modifiable process that can be tuned by the manipulation of specific deubiquitinating enzymes or proteasome activity. This possibility further provides opportunities for therapeutic interventions (Lipkowitz and Weissman, 2011; Crawford et al., 2011). Relevant to OPMD, proteasome activity is reduced during muscle aging (Combaret et al., 2009; Lee et al., 1999; Ferrington et al., 2005) and perhaps, consequently, leads to accumulation of altered proteins. Concordantly, the expression of many aggregation-prone proteins was found to be deregulated in OPMD as well as in other protein aggregation disorders (Anvar et al., 2011a; Corbeil-Girard et al., 2005). Our analysis revealed that the core subunit of the proteasome was consistently down-regulated in OPMD (Anvar et al., 2011a). Additionally, we have shown that expPABPN1 expression in myotubes leads to down-regulation of proteasome-encoding genes and affects the accumulation of expPABPN1 protein (Raz et al., 2011). In turn, manipulation of proteasome activity also affects the accumulation and aggregation of expPABPN1 (Anvar et al., 2011a; Raz et al., 2011). In spite of this prominent link between proteasome activity, expPABPN1 accumulation and INI formation, this process is not specific to muscle cells (Abu-Baker et al., 2003). Since the onset of OPMD coincides with proteasomal down-regulation in ageing muscle, it is possible that the decline in proteasome activity during muscle aging triggers or accelerates expPABPN1 accumulation. Subsequently, in OPMD, aggregation of mutant PABPN1 leads to extensive proteasome down-regulation and entrapment of proteasomal proteins in INIs. This feed forward model along with the onset of skeletal muscle ageing could explain the muscle-specific and INI formation in OPMD (**Figure 2**). Notably, decrease in skeletal muscle performance, as measured by muscle strength, strongly correlates with chronological ageing (Beenakker et al., 2010). Loss of muscle function during ageing is regulated by numerous genetic and environmental factors (Roth et al., 2002) which may explain the differences in muscle performance among individuals (Kostek and Delmonico, 2011). Ageing associated physiological changes can be accompanied by an increased susceptibility to degenerative disorders (Kirkwood and Austad, 2000). Although in most tissues ageing is marked by a progressive decline of cellular functions starting at mid-life (Kirkwood, 2005; Lexell et al., 1988; Lindle et al., 1997; Sahin and Depinho, 2010), the rate of functional changes is tissue-specific (Kirkwood and Austad, 2000).

We found substantial similarities in transcriptional changes between muscle ageing and OPMD. The most striking finding, based on the analysis of expression profiles, was the significant decline in *PABPN1* expression during the first half of the fifth decade. Since changes in skeletal muscle performance commence at the fifth decade (Lindle et al., 1997; Roth et al., 2002) our results suggest a correlation between *PABPN1* expression and the onset of muscle ageing. Moreover, among controls, *PABPN1* expression in females was significantly lower than in males. This observation is in agreement with previous studies indicating that ageing-associated changes in muscle strength are more pronounced in females (Kent-Braun et al., 2002; Roth et al., 2002). Concordantly, the OPMD prevalence of the Uruguayan population is estimated to be higher in females (Medici et al., 1997). Thus, the *PABPN1* expression profile could additionally mark gender-associated decline in muscle performance. Together, the progressive decline in *PABPN1* expression during muscle ageing and the accelerated reduction of its expression in OPMD indicate a strong correlation with muscle weakness. The early mid-life onset of *PABPN1* down-regulation (as compared to that of frontal cortex brain tissues (Lu et al., 2004) with the onset of 85 years; and *Rectus Abdominis* (Zahn et al., 2006) tissues with unchanged expression) suggests temporal-spatial specific-

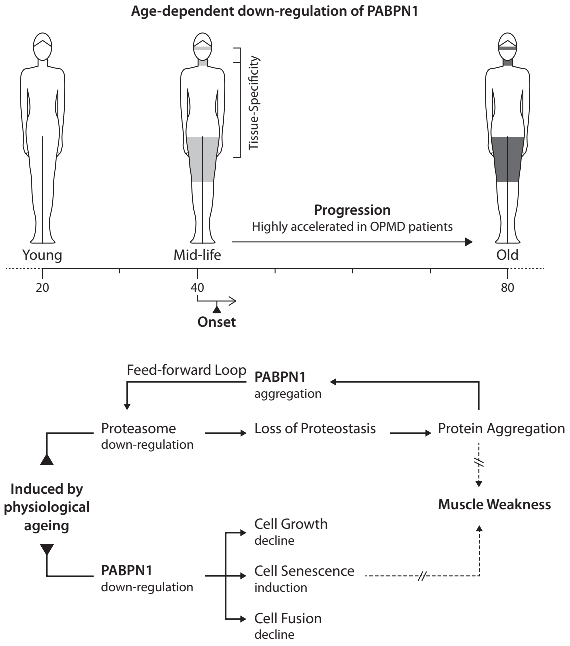


Figure 2 – A model for molecular mechanisms involved in OPMD pathology. In muscles, age-associated proteasome down-regulation triggers expPABPN1 protein accumulation. Subsequently, elevated expPABPN1 aggregation leads to proteasome deregulation during disease onset. This feed forward loop and the onset of muscle ageing leads to loss of proteostasis and INI formation. As part of ageing-related transcriptional changes, there is a significant reduction in the expression of PABPN1. This age-associated decline is accelerated in OPMD patients. In cell cultures, reduced expression of PABPN1 during ageing of skeletal muscles leads to progressive cell senescence and defects in cell fusion and growth. The effect on the expression of muscle contraction genes highly depends on the level of PABPN1 expression. The decline in PABPN1 expression may partially explain the progressive decline in muscle performance during ageing and accelerated muscle weakness in OPMD patients.

ity. However, some reports have indicated mental retardation, cognitive impairment, spinal cord involvement, and dementia in some OPMD patients (Millefiorini and Filippini, 1967; Sarkar et al., 1995; Blumen et al., 2009; Linoli et al., 1991; Mizoi et al., 2011; Dubbioso et al., 2011). Thus, it would be interesting to assess PABPN1 expression in respect to the central nervous system.

Age-dependent progressive decline of PABPN1 expression and loss of muscle function suggests that PABPN1 may play a role in ageing of skeletal muscles (see Chapter **three**). PABPN1 expression in OPMD patients is only 30% of that found in young healthy controls. In immortalized human myoblast cultures, this expression level leads to progressive cellular defects including reduced cell growth and fusion and induced cell senescence. Heterochromatic foci (HF), the hallmark of cellular senescence (Spector and Gasser, 2003), could be observed in cells with 70% PABPN1 down-regulation. Notably, PABPN1 expression was undetectable in nuclei with HF. We suggest that the effect of PABPN1 down-regulation on cellular senescence is more pronounced in non-mitotic cells as they exhibit a three-fold higher amount of cells with HF. Myotube cultures from OPMD muscles also show premature senescence and reduced cell fusion (Perie et al., 2006). Relevant to reduced muscle performance, the expression of muscle contraction genes highly depend on PABPN1 expression level. Recently, we showed that increased PABPN1 protein accumulation in muscle cells results in a reduced amount of the soluble and functional protein (Raz et al., 2011). Since PABPN1 regulates mRNA stability it is expected that decline in functional PABPN1 would have a broad effect on cellular functions as demonstrated here and by Apponi et al. (Apponi et al., 2010). Together, for the first time, our data indicates the progressive response of muscle cell function to the level of PABPN1 in a spatial-temporal manner, highlighting PABPN1 role as a key regulator of muscle ageing.

Ageing cells exhibit distinctive features ranging from the accumulation of damaged macromolecules to changes in nuclear architecture (Campisi and Vijg, 2009; Oberdoerffer and Sinclair, 2007). In particular, it has been suggested that ageing and age-related disorders are strongly associated with mechanisms that control chromatin structure through DNA methylation, RNA interference, histone variants, and post-translational modifications (Oberdoerffer and Sinclair, 2007; Campisi and Vijg, 2009; Rakyan et al., 2010; Teschendorff et al., 2010; Esteller, 2007; Pogribny et al., 2006; Tryndyak et al., 2006; Ronn et al., 2008; Tohgi et al., 1999; Martin, 2009; Rando and Chang, 2012). Nuclear chromatin is associated with processes that mediate DNA replication and transcription (Trinkle-Mulcahy and Lamond, 2007). Markedly, gene transcription is strongly modulated by its relative position within the nucleus (Sexton et al., 2007). This suggests that disruption of the positioning of the chromatin at the nuclear envelop can affect the regulation of gene expression (Akhtar and Gasser, 2007). Furthermore, DNA and chromatin modifications are recognized as both responsive and effectors of the ageing process (Martin, 2009; Rando and Chang, 2012). Therefore, the spatial distribution of genes across the nuclear envelop can significantly contribute to the transcriptional control. Aged cells, in particular, show several changes on their chromatin and nuclear envelop structure that contribute to the lineage and tissue-specific gene expression (Krishnamurthy et al., 2004; Rando and Chang, 2012). It will be interesting to investigate the possibility in which epigenetic changes play a role in mechanisms that underlie the onset and progression of OPMD. Identification of possible epigenetic factors that may be functionally associated with the OPMD phenotype can provide insights on the relationship between the genome and environment. This would potentially lead to a better understanding and characterization of the severity of symptoms in OPMD patients.

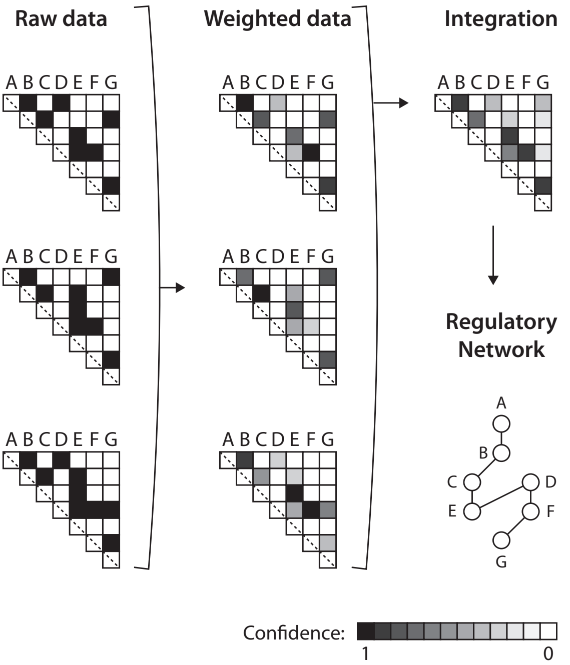
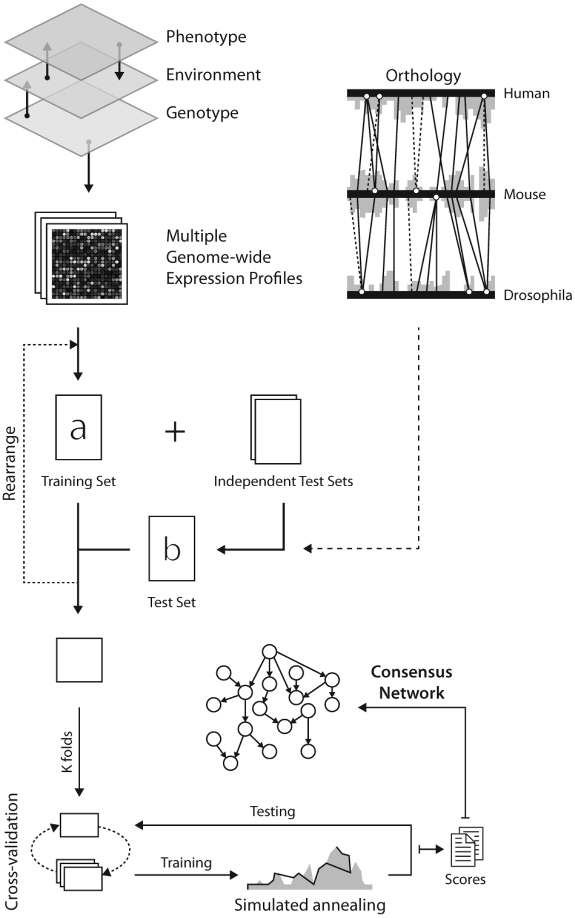


Figure 3 – Probabilistic network integration. Datasets from multiple sets of independent experiments on different species are individually tested and optimised for their association with a given phenotype. Various statistical approaches can be used to infer confidence weight for any given intraspecies regulatory relationship. These weights can then be used to integrate network structures across species. Graphical networks can be derived from the final weighted matrix after applying a confidence threshold.

PABPN1 is involved in pre-mRNA polyadenylation, where it stimulates poly(A) polymerase and regulates poly(A) tail length and RNA stability (Lemay et al., 2010; Kuhn et al., 2009). It is now widely accepted that alternative processing of pre-mRNA can result in structural variation and differing function of encoded proteins (Moore and Silver, 2008; Birzele et al., 2008), as well as regulation of gene expression. Elongation or shortening of the 3' un-translated region (UTR), as a consequence of alternative polyadenylation, can lead to changes in binding of miRNAs and, therefore, differential regulation of mRNAs. Considering the role of PABPN1 in regulating the poly(A) tail and initial indications regarding a widespread discordance between deregulated transcripts of genes, it is crucial to pursue such investigation using next-generation sequencing. Moreover, mechanisms that regulate the 3' UTR are controlled in a tissue-specific manner. There-

Figure 4 – Schematic overview of the Dandelion algorithm for disease network analysis. The Dandelion algorithm involves three recurring stages of training and an independent testing regime with the use of multiple datasets derived from different species. In the first step, disease modules are defined based on prior knowledge. The next step involves reiterative selection of one species for which the gene regulatory network is constructed while others are left aside for independent testing and validation of the learnt disease networks. For the construction of an intraspecies disease network, the dataset is divided into k-folds, using cross-validation. Subsequent, regulatory relationships between gene transcripts are learnt using a Bayesian network methodology based upon simulated annealing optimization of the network Bayes Information Criterion (BIC) score. After applying confidence thresholds on relationships between genes, the disease network is translated to the expected interspecies disease network. This is achieved by the use of the cross-validation and network optimization procedure. The algorithm searches through the relationships found in the training dataset to find the best fit for the interspecies representation of the disease network. These networks are then integrated by removing all the links with low confidence score across species.



fore, it is essential to pinpoint how ageing-associated decline of soluble PABPN1 and induced aggregation of the mutant PABPN1 may lead to tissue-specific changes in poly(A) site usage. In addition, diversification of RNA, and consequently protein function and structure, is regulated through processes of which alternative splicing plays a central role. In particular, skeletal muscle is reported as one of the tissues with the highest rate of alternative splicing (Pan et al., 2008; Wang et al., 2008; Castle et al., 2008). It is not surprising that genetic mutations may lead to deregulation of this process and consequently cause a widespread transcriptional changes (Cooper et al., 2009; Wang and Cooper, 2007; Tazi et al., 2009). Thus, it is important to pursue the possibility that alternative splicing is differentially regulated in OPMD patients and model systems.

The work, presented in this thesis, strongly highlights the fascinating nature and value of interdisciplinary studies. We have shown that the concept of a universality of biological processes in the light of evolutionary mechanisms and common functional processes can lead to novel discoveries. Engaging in the study of a variety of organisms or biological behaviours, looking for shared molecular features in a rare disease such as OPMD, enabled us to uncover insights on a broader spectrum of conditions and phenomena such as ageing of skeletal muscles and protein aggregation disorders.

The inner workings of complex biological networks

Functional interdependencies and the modular nature of cell's molecular components imply the indispensable role of network-based approaches to human diseases. It is widely established that these dependencies revealed by regulatory networks can provide valuable information regarding underlying biological processes (Avery and Wasserman, 1992; Tong et al., 2004; Costanzo et al., 2010; St Onge et al., 2007; Schuldiner et al., 2008; Collins et al., 2007; Segre et al., 2005; Drees et al., 2005; Guarente, 1993; Hartman et al., 2001; Jonikas et al., 2009). Despite the generation of vast quantities of data by high-throughput technologies, biological data are usually sparse, noisy and ambiguous, limited in number of samples, and high-dimensional. Thus, integration of data and genomic information from human and various model systems can ultimately provide a better indication of common molecular mechanisms that underlie a given phenotype. However, the presence of noise and technical artefacts specific to model systems usually leads to limited overlap between results obtained in cross-species comparison (Lu et al., 2009; Zhou and Gibson, 2004; Oliva et al., 2005; Blake et al., 2003; Jelier et al., 2008). Additionally, integrative approaches are far from trivial and are complicated due to our limited knowledge of true protein orthologues, transcript variants coding for proteins with similar function, and evolutionary conservation of biological processes. These bottlenecks further require fine tuning and optimization of the integration strategy. Another aspect of complexity arises from the generation of large-scale networks (having thousands of nodes and millions of possible interactions), owing to limited computational power and intelligent algorithms for scalability and reducing dimensionality (Venkatesan et al., 2009; Barabasi et al., 2011). Markedly, such stochastic systems require a probabilistic approach at the core for modelling regulatory networks.

We first established, in Chapter **four**, a way in which gene networks that are highly informative for determining "muscle differentiation" can be robustly identified from multiple independent datasets with increasing level of complexity and stochasticity (Anvar et al., 2010). We showed that the proper use of a modelling strategy in combination with multiple datasets leads to the construction of gene networks that can explain the myogenesis-related genes significantly better than those that have less involvement in myogenesis. This approach resulted in networks that were consistently more parsimonious to myogenesis-related genes. Moreover, these models provide the robust prediction of biological outcome and expression profiles. Establishing a strategy which can accommodate the integration of multiple datasets enables the possibility of overcoming the limitations of cross-species integrative studies. Such exploitation would lead to more robust regulatory mechanisms to be identified and predictions to be made across various platforms and organisms (**Figure 3** and **Figure 4**). In Chapter **five**, we showed that the integration and analysis of microarray datasets from various species increase the robustness of the constructed networks and the predictive accuracy of the disease state (Anvar et al., 2011b). We also demonstrated that the interspecies translation of these networks helps to avoid overfitting. In addition, this approach provides a state-of-the-art model-driven selection of transcript isoforms that are most likely to be coding for orthologous proteins. Notably, another fascinating application of this strategy would be the identification of alternative splicing events and their regulators (Zhang et al., 2010). These powerful features are essential for understanding the phenotypic implications of such strong relationships as part of evaluating the conservation and dynamics of interspecies disease networks. Moreover, the high level of specificity and sensitivity of these models enables the prioritization of candidate regulators of the disease molecular mechanisms to be studied in follow-up validation experiments. In particular, it is crucial to carry out additional experiments to investigate the tissue-specificity of the network (Reverter et al., 2008; Lage et al., 2010) and the functional relevance of encoded proteins dysregulation to the disease pathology. This can also be achieved by re-constructing tissue- or cell-specific sub-networks from the model by integrating a variety of tissue-specific data sources (Jerby et al., 2010; Kirouac et al., 2010).

Our approach for Bayesian modelling of datasets on a similar phenotype from different model systems and patients is unique. Several approaches have been described to avoid overfitting and increase the robustness of Bayesian networks. For example, informative priors derived from protein-protein interaction (PPI) data or from the literature have been used to generate more stable and biologically meaningful networks (Segal et al., 2003; Pe'er et al., 2002; Steele et al., 2009; Jansen et al., 2003). While these methods obviously bias the results towards well-known regulatory interactions and are less likely to detect novel relationships (Sprinzak et al., 2003; Joyce and Palsson, 2006), they may ultimately be combined with our modelling approach to obtain regulatory networks with a more straightforward biological interpretation.

Our method was applied to an *a priori* defined gene module coding for a well-known biological structure, the proteasome. Several studies in *S. cerevisiae* (Zhang et al., 2005; Tanay et al., 2004; Luscombe et al., 2004; Han et al., 2004) have demonstrated the value of an integrative modelling approach providing modularized interaction networks without prior assumptions. Zhang et al. (Zhang et al., 2005), for instance, took an approach in which they integrated a number of different available data sources, from PPIs to sequence homology and gene co-expression, while Tanay et al. (Tanay et al., 2004) and others (Luscombe et al., 2004; Han et al., 2004) expanded on the statistical analysis of network properties and identified modules within the network structure. The performance of

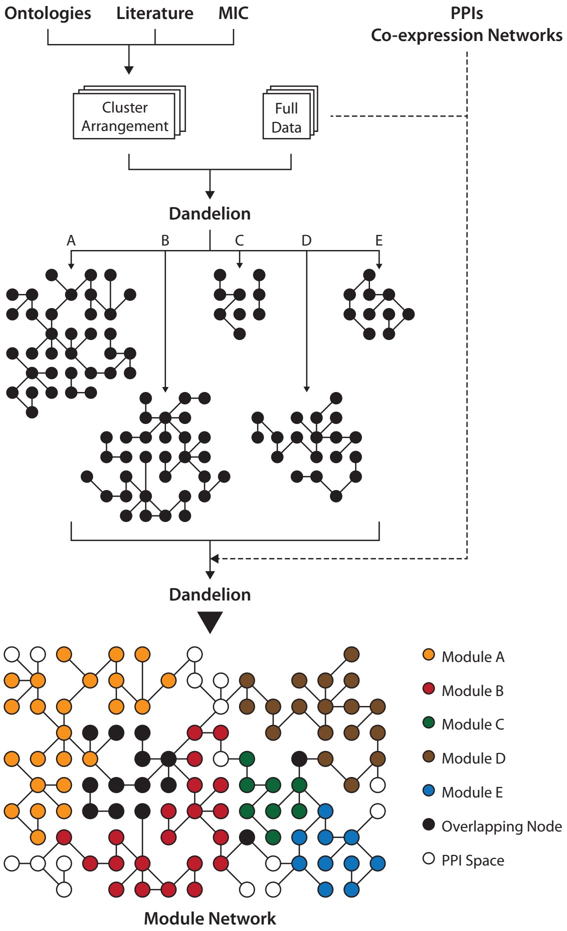


Figure 5 – Schematic illustration of Dandelion module networks. In a bottom-up approach, gene modules are curated on the basis of their literature-aided and cross-species association or according to predefined ontologies. The algorithm involves three recurring phases of training and an independent testing regime with the use of multiple datasets from different platforms, experiments, or organisms. First, consensus networks are constructed for individual modules using our previously described Dandelion algorithm (Chapter **five**). These sub-networks are then overlaid based on common nodes and relationships within different network structures. Finally, using protein-protein interaction databases or associations in co-expression networks, the Dandelion algorithm attempts to assemble and optimise the full module network by adding relationships and nodes to interlink sub-networks. Additionally, the Dandelion algorithm would allow for novel nodes and relationships to be added to the global module network structure. The growth of module networks is constrained on the overall improvement of the network performance.

these models depends on the availability of high quantities of samples and may be prone to overfitting due to the presence of noise and other model-specific artefacts. Therefore, a combination of their approach with our interspecies translation may enable the discovery of larger gene regulatory networks with multiple gene modules and connections between them.

Understanding the dynamics of network structure is essential for determining causal interdependencies as well as characterization of network modularity and gene spatial properties. In model organisms, it has been shown that hub proteins are tend to be encoded by essential genes (Jeong et al., 2001) which are highly conserved (Fraser et al., 2002; Eisenberg and Levanon, 2003; Saeed and Deane, 2006). Identification of essential genes is important for discovery of sub-networks that are associated with a disease phenotype, owing to the disease-related genes being located in the network-based vicinity of the hub nodes (Goh et al., 2007; Feldman et al., 2008). The importance of nodes to the network can be estimated using the 'betweenness centrality' measure (Yu et al., 2007) which gives some additional insights on topology, information flow, and the stability of a network (Han, 2008). Topological properties of disease networks reveal clouds of densely interconnected nodes that can be used for gene module prediction (Girvan and Newman, 2002; Palla et al., 2005; Ahn et al., 2010; Enright et al., 2002). In addition to network topology, functional characterisation of subnetworks can improve in describing mechanisms that give rise to a specific phenotype.

Here, I discuss a strategy to tackle some challenges in bridging the gap between multi-layers of biological data. In a study presented in Chapter **five**, we developed a novel algorithm for constructing interspecies disease networks that provide an assumption-free and model-driven selection of the most important transcript isoforms across species (Anvar et al., 2011b). We achieved this by use of prior knowledge on pathways that are disease-associated. This was owing to the fact that, on a genome-wide scale, searching the space of possible networks via single-arc changes is not realistic and computationally expensive. One of the possible strategies for reducing the high dimensionality is the use of statistical algorithms such as ridge and LASSO regression (Friedman et al., 2008; Tibshirani, 1996). These algorithms apply a penalty for complex models that may be tuned by cross-validation. However, this would mean that the same dataset is

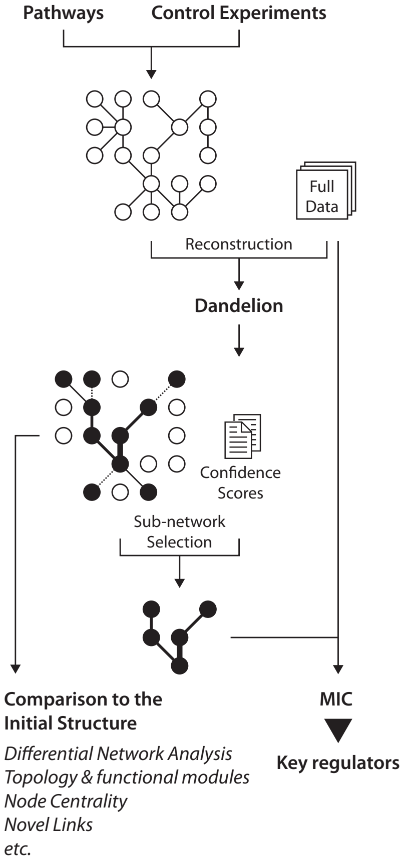


Figure 6 – A model for network reconstruction and evaluation. Known networks, produced based on control experiments or molecular pathways, can be reconstructed using the Dandelion algorithm. Reconstructed networks differ in respect to nodes being incorporated (depicted in black) within the network structure. Moreover, the comparative analysis can unravel changes in the dynamics of such regulatory networks under different phenotype or experimental settings. Relationships with the strongest weights (depicted by the line thickness), on the basis of their confidence score, can be evaluated using the maximal information coefficient (MIC) measure. These strategies could uncover key regulators of a given pathway under specific phenotype or experimental setting.

used, in two steps, for generating the network space and construction of disease network which would lead to overfitting and other biases. Another alternative is based on treating functional modules as blocks of interconnected nodes which can be assembled together by the use of overlaying nodes and relationships. In conjunction with assembling overlaying nodes and modules, additional links are added to the network for global optimization of the inter-module relationships (**Figure 5**). This can be reached by simple hill-climbing, greedy algorithms or more sophisticated simulated annealing and MCMC (Markov Chain Monte Carlo) searching methods. Within the optimization step, evidences from PPI networks can be used for confidence assessment. In addition to the utility of PPI networks, reconstruction of known functional pathways, or those produced by alternative models on control datasets, can be combined with allowing for novel relationships (Battle et al., 2010) (**Figure 6**). This strategy potentially can help automating the process of optimization and confidence assessment. Moreover, the maximal information coefficient (Reshef et al., 2011) can be integrated to assess the functional association for relationships in the vicinity of the essential nodes.

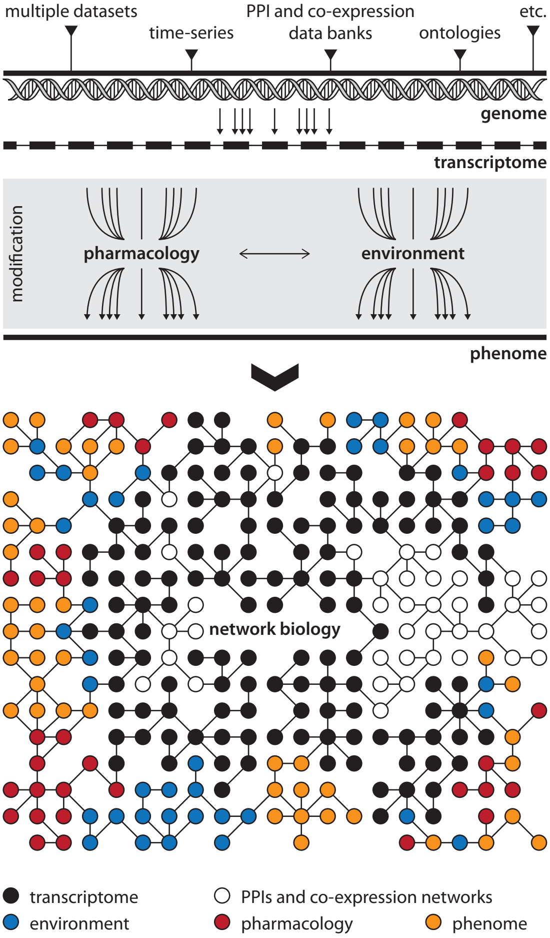


Figure 7 – Network medicine, linking across multi-layers of biological data. Datasets from different organisms or platforms can be combined for enhanced identification of interspecies (inter-platform) networks. Time-series datasets can provide information on the dynamics of biological networks while protein-protein interaction and co-expression networks can be used for optimization and scaling. Networks constructed on transcriptome data are linked to networks related to pharmacology, phenomics, and environment. Genomic information, reflected in transcriptome, are interlinked and translated to diverse sets of phenotypes through environmental factors. Likewise, these multi-layers of densely interconnected regulatory relationships are represented through a framework of pharmacological entities.

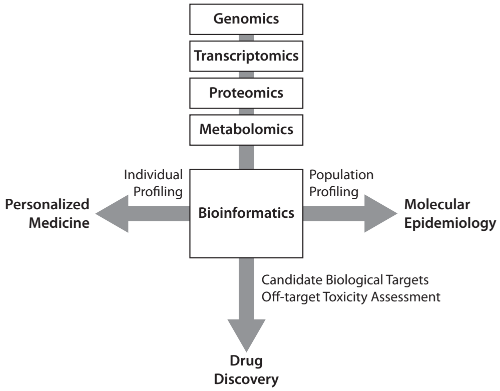
Finally, the evolution of these network properties over time would provide a crucial framework for better understanding the causal relationship and dynamics of gene regulatory networks in the context of human diseases. Thus, I believe that robust and unbiased construction and analysis of the interspecies networks for rare or complex human diseases can lead to novel discovery and identification of key regulators. The result of such exploration can ultimately offer potential targets for therapeutic interventions and drug developments. In the last section, I will discuss a few strategies that, in my view, can be pursued to enhance data integration and the ideal utility of network-based approaches on a larger scale. This would consequently provide a disease-oriented global view of genomics, transcriptomics, proteomics, and newly defined field of phenomics. The term "phenomics" is derived from the word 'phenome' which was first introduced by Michael Soule (Soule, 1967).

From systems biology to personalized medicine

In recent years, the studies of human diseases have changed significantly, owing to advancements in the field of systems biology and high-throughput technologies. It is widely believed that the integration of genomics, transcriptomics, and large-scale phenotyping has major potential for novel discoveries in network biology of multi-layered human disorders (Bilder et al., 2009; Freimer and Sabatti, 2003; Schilling et al., 1999; Searls, 2005). Phenotypic variations are determined by a complex network of genetic and environmental interactions. In the last two decades, significant efforts have been made on genomic and transcriptomic studies. Now we have reached the time to invest special efforts in the field of phenomics which still requires our careful attention. This is due to the lack of standardisation and data available. Although limited phenotyping efforts for obvious disease-related features seems to be sufficient in most cases, extensive and global phenotyping can pave the way for better standardization of phenotypic information and mechanistic understanding of context-oriented genetic and environmental interactions. This would lead to the discovery of novel dependencies between genomics, transcriptomics, and phenomics data. Additionally, this combinatory framework provides an information-rich model that can distinctly characterise the correlation or causal relationships and account for different sources of variation (Houle et al., 2010). The importance of the integrative approaches is evident from experiments carried out in yeast that demonstrate a substantial growth from lethal or disease causing single-gene deletions (34%) to those that occur in conjunction with at least one environmental condition (97%) (Hopkins, 2008). Nevertheless, the design of an integrative strategy needs to be addressed with precision and care as navigating such data is extremely challenging. For instance, one of the basic information losses is that phenotypic data is often treated as a discretised entity whereas the most vital piece of information lays in the relative and continuous changes of phenotypic information, phenomena which is now well-established for other data-types such as transcriptome.

Having established the modular network structure, the next step in exploring the interplay between transcriptomics and phenotypic states of human diseases is to determine the environmental factors through which these networks are regulated in a full range of spatial and temporal scale. Adequate combination of prior knowledge (Ochs, 2010) can further provide confirmatory insights on data-flow and ordering of causal relationships across multi-layers of biological data (**Figure 7**). Yet, the use of prior knowledge-centric approaches needs to be avoided to minimize the biases that can be introduced by such techniques. An intriguing possibility of such methods is that the construction of multifaceted biological networks may provide insights on efficacy and off-target toxicity of drugs in a phenotype-centric and tissue-specific manner, some of which can be determined by the analysis of such network structure (Albert et al., 2000; Kitano, 2007). Likewise, special efforts in modelling the dynamics of metabolic responses in different tissues can provide valuable insights into the effects of drugs and diseases (**Figure 8**). Another intriguing benefit of engaging in metabolomics studies is the possibility of linking different levels of biological organization (genomics, transcriptomics, proteomics, etc.), owing to their differing operational behaviour (Holmes et al., 2008a; Holmes et al., 2008b; Nicholson and Wilson, 2003). An extensive review by Hopkins (Hopkins, 2008) provides valuable information on the usability of biological networks in drug discovery along with a brief outlook on future prospects. While some of these advancements seem farfetched and years in the future, a few preliminary developments can be pursued that provides a new basis for a global infrastructure of network medicine. For in-

Figure 8 – Applications of multifaceted omics research. Special efforts in combined analysis across multi-layers of biological data provide an infrastructure in which bioinformatics can play a central role. Naturally, the main applications of omics research can be divided into three fields of personalized medicine, drug discovery, and molecular epidemiology. Profiling of individuals can provide an enhanced framework for better therapeutic interventions. The utility of this strategy is to comprehend patients' susceptibility to diseases or alter therapies on the basis of their response to different medicine. Molecular epidemiology studies can be enhanced by looking for common patterns in profiles within a population. This would allow for the identification of biomarkers, susceptibilities of specific populations to diseases, and health screening programmes. Finally, these studies can lead to uncovering new biological targets for drug discovery.



stance, the adoption of methods that deal with dynamics of these networks, in a spatial-temporal manner, can act as a cornerstone for robust integration of pharmaceutical data and chemical interactions. This combinatory strategy provides a valuable framework for drug discovery and personalized therapeutic interventions. Notably, recent approaches for simple characterization of the network topology had made a remarkable contribution in developing strategies for prioritization and combination of drug targets (Gerber et al., 2008; Potapov et al., 2008; Wunderlich and Mirny, 2006). Mining biomedical and biochemical literature in conjunction with ontologies (such as KEGG and GO) are also well-explored to better determine the efficacy of drug development (Yildirim et al., 2007; Ji et al., 2007; Spiro et al., 2008; Gunther et al., 2008). Bayesian approaches can bridge between these different sources of information and provide a global network infrastructure in which transcriptome data, environmental factors, and phenotypic information can come together to provide a predictive and model-driven framework for assessing the clusters of chemical networks and pharmacology data (**Figure 7**). To conclude, the context- and case-specific identification of the optimal point of interaction between molecules for drug discovery is the future of systems biology applications in the field of personalized medicine. In order to achieve this ambition, novel and integrative advancements are needed to better understand the global organisation of networks in the study of human genetic disorders.

Reference List

Abu-Baker,A., Messaed,C., Laganiere,J., Gaspar,C., Brais,B., and Rouleau,G.A. (2003). Involvement of the ubiquitin-proteasome pathway and molecular chaperones in oculopharyngeal muscular dystrophy. Hum. Mol. Genet *12*, 2609-2623.

Ahn,Y.Y., Bagrow,J.P., and Lehmann,S. (2010). Link communities reveal multiscale complexity in networks. Nature *466*, 761-764.

Akhtar,A. and Gasser,S.M. (2007). The nuclear envelope and transcriptional control. Nat Rev Genet. *8*, 507-517.

Albert,R., Jeong,H., and Barabasi,A.L. (2000). Error and attack tolerance of complex networks. Nature *406*, 378-382.

Anvar,S.Y., 't Hoen,P.A., and Tucker,A. (2010). The identification of informative genes from multiple datasets with increasing complexity. BMC. Bioinformatics. *11*, 32.

Anvar,S.Y., 't Hoen,P.A., Venema,A., van der Sluijs,B., van,E.B., Snoeck,M., Vissing,J., Trollet,C., Dickson,G., Chartier,A., Simonelig,M., van Ommen,G.J., van der Maarel,S.M., and Raz,V. (2011a). Deregulation of the ubiquitin-proteasome system is the predominant molecular pathology in OPMD animal models and patients. Skelet. Muscle *1*, 15.

Anvar,S.Y., Tucker,A., Vinciotti,V., Venema,A., van Ommen,G.J., van der Maarel,S.M., Raz,V., and 't Hoen,P.A. (2011b). Interspecies translation of disease networks increases robustness and predictive accuracy. PLoS. Comput. Biol. *7*, e1002258.

Apponi,L.H., Leung,S.W., Williams,K.R., Valentini,S.R., Corbett,A.H., and Pavlath,G.K. (2010). Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. Hum. Mol. Genet. *19*, 1058-1065.

Avery,L. and Wasserman,S. (1992). Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet. *8*, 312-316.

Barabasi,A.L., Gulbahce,N., and Loscalzo,J. (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet. *12*, 56-68.

Battle,A., Jonikas,M.C., Walter,P., Weissman,J.S., and Koller,D. (2010). Automated identification of pathways from quantitative genetic interaction data. Mol. Syst. Biol. *6*, 379.

Beenakker,K.G., Ling,C.H., Meskers,C.G., de Craen,A.J., Stijnen,T., Westendorp,R.G., and Maier,A.B. (2010). Patterns of muscle strength loss with age in the general population and patients with a chronic inflammatory state. Ageing Res. Rev *9*, 431-436.

Ben-Zvi,A., Miller,E.A., and Morimoto,R.I. (2009). Collapse of proteostasis represents an early molecular event in Caenorhabditis elegans aging. Proc. Natl. Acad. Sci. U. S. A *106*, 14914-14919.

Ben-Zvi,A.P. and Goloubinoff,P. (2002). Proteinaceous infectious behavior in non-pathogenic proteins is controlled by molecular chaperones. J. Biol. Chem. *277*, 49422-49427.

Bilder,R.M., Sabb,F.W., Cannon,T.D., London,E.D., Jentsch,J.D., Parker,D.S., Poldrack,R.A., Evans,C., and Freimer,N.B. (2009). Phenomics: the systematic study of phenotypes on a genome-wide scale. Neuroscience *164*, 30-42.

Birzele,F., Csaba,G., and Zimmer,R. (2008). Alternative splicing and protein structure evolution. Nucleic Acids Res. *36*, 550-558.

Blake,W.J., KAErn,M., Cantor,C.R., and Collins,J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633-637.

Blumen,S.C., Bouchard,J.P., Brais,B., Carasso,R.L., Paleacu,D., Drory,V.E., Chantal,S., Blumen,N., and Braverman,I. (2009). Cognitive impairment and reduced life span of oculopharyngeal muscular dystrophy homozygotes. Neurology *73*, 596-601.

Bodine,S.C., Latres,E., Baumhueter,S., Lai,V.K., Nunez,L., Clarke,B.A., Poueymirou,W.T., Panaro,F.J., Na,E., Dharmarajan,K., Pan,Z.Q., Valenzuela,D.M., DeChiara,T.M., Stitt,T.N., Yancopoulos,G.D., and Glass,D.J. (2001). Identification of ubiquitin ligases required for skeletal muscle atrophy. Science *294*, 1704-1708.

Calado,A., Kutay,U., Kuhn,U., Wahle,E., and Carmo-Fonseca,M. (2000). Deciphering the cellular pathway for transport of poly(A)-binding protein II. RNA. *6*, 245-256.

Campisi,J. and Vijg,J. (2009). Does damage to DNA and other macromolecules play a role in aging? If so, how? J. Gerontol. A Biol. Sci. Med. Sci. *64*, 175-178.

Cao,P.R., Kim,H.J., and Lecker,S.H. (2005). Ubiquitin-protein ligases in muscle wasting. Int. J. Biochem. Cell Biol. *37*, 2088-2097.

Castle,J.C., Zhang,C., Shah,J.K., Kulkarni,A.V., Kalsotra,A., Cooper,T.A., and Johnson,J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet. *40*, 1416-1425.

Chartier,A., Benoit,B., and Simonelig,M. (2006). A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. EMBO J *25*, 2253-2262.

Ciechanover,A. and Brundin,P. (2003). The ubiquitin proteasome system in neurodegenerative diseases: sometimes the chicken, sometimes the egg. Neuron *40*, 427-446.

Collins,S.R., Miller,K.M., Maas,N.L., Roguev,A., Fillingham,J., Chu,C.S., Schuldiner,M., Gebbia,M., Recht,J., Shales,M., Ding,H., Xu,H., Han,J., Ingvarsdottir,K., Cheng,B., Andrews,B., Boone,C., Berger,S.L., Hieter,P., Zhang,Z., Brown,G.W., Ingles,C.J., Emili,A., Allis,C.D., Toczyski,D.P., Weissman,J.S., Greenblatt,J.F., and Krogan,N.J. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature *446*, 806-810.

Combaret,L., Dardevet,D., Bechet,D., Taillandier,D., Mosoni,L., and Attaix,D. (2009). Skeletal muscle proteolysis in aging. Curr. Opin. Clin. Nutr. Metab Care *12*, 37-41.

Cooper,T.A., Wan,L., and Dreyfuss,G. (2009). RNA and disease. Cell *136*, 777-793.

Corbeil-Girard,L.P., Klein,A.F., Sasseville,A.M., Lavoie,H., Dicaire,M.J., Saint-Denis,A., Page,M., Duranceau,A., Codere,F., Bouchard,J.P., Karpati,G., Rouleau,G.A., Massie,B., Langelier,Y., and Brais,B. (2005). PABPN1 overexpression leads to upregulation of genes encoding nuclear proteins that are sequestered in oculopharyngeal muscular dystrophy nuclear inclusions. Neurobiol. Dis. *18*, 551-567.

Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L., Toufighi,K., Mostafavi,S., Prinz,J., St Onge,R.P., VanderSluis,B., Makhnevych,T., Vizeacoumar,F.J., Alizadeh,S., Bahr,S., Brost,R.L., Chen,Y., Cokol,M., Deshpande,R., Li,Z., Lin,Z.Y., Liang,W., Marback,M., Paw,J., San Luis,B.J., Shuteriqi,E., Tong,A.H., van,D.N., Wallace,I.M., Whitney,J.A., Weirauch,M.T., Zhong,G., Zhu,H., Houry,W.A., Brudno,M., Ragibizadeh,S., Papp,B., Pal,C., Roth,F.P., Giaever,G., Nislow,C., Troyanskaya,O.G., Bussey,H., Bader,G.D., Gingras,A.C., Morris,Q.D., Kim,P.M., Kaiser,C.A., Myers,C.L., Andrews,B.J., and Boone,C. (2010). The genetic landscape of a cell. Science *327*, 425-431.

Crawford,L.J., Walker,B., and Irvine,A.E. (2011). Proteasome inhibitors in cancer therapy. J. Cell Commun. Signal. *5*, 101-110.

Davies,J.E., Sarkar,S., and Rubinsztein,D.C. (2006). Trehalose reduces aggregate formation and delays pathology in a transgenic mouse model of oculopharyngeal muscular dystrophy. Hum. Mol. Genet *15*, 23-31.

Deshaies,R.J. and Joazeiro,C.A. (2009). RING domain E3 ubiquitin ligases. Annu. Rev Biochem. *78*, 399-434.

Drees,B.L., Thorsson,V., Carter,G.W., Rives,A.W., Raymond,M.Z., Avila-Campillo,I., Shannon,P., and Galitski,T. (2005). Derivation of genetic interaction networks from quantitative phenotype data. Genome Biol. *6*, R38.

Dubbioso,R., Moretta,P., Manganelli,F., Fiorillo,C., Iodice,R., Trojano,L., and Santoro,L. (2011). Executive functions are impaired in heterozygote patients with oculopharyngeal muscular dystrophy. J. Neurol.

Eisenberg,E. and Levanon,E.Y. (2003). Preferential attachment in the protein network evolution. Phys. Rev Lett. *91*, 138701.

Emanuele,M.J., Elia,A.E., Xu,Q., Thoma,C.R., Izhar,L., Leng,Y., Guo,A., Chen,Y.N., Rush,J., Hsu,P.W., Yen,H.C., and Elledge,S.J. (2011). Global identification of modular cullin-RING ligase substrates. Cell *147*, 459-474.

Enright,A.J., Van,D.S., and Ouzounis,C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575-1584.

Esteller,M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet. *8*, 286-298.

Feldman,I., Rzhetsky,A., and Vitkup,D. (2008). Network properties of genes harboring inherited disease mutations. Proc. Natl. Acad. Sci. U. S. A *105*, 4323-4328.

Ferrington,D.A., Husom,A.D., and Thompson,L.V. (2005). Altered proteasome structure, function, and oxidation in aged muscle. FASEB J. *19*, 644-646.

Fraser,H.B., Hirsh,A.E., Steinmetz,L.M., Scharfe,C., and Feldman,M.W. (2002). Evolutionary rate in the protein interaction network. Science *296*, 750-752.

Freimer,N. and Sabatti,C. (2003). The human phenome project. Nat Genet. *34*, 15-21.

Friedman,J., Hastie,T., and Tibshirani,R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics. *9*, 432-441.

Gerber,S., Assmus,H., Bakker,B., and Klipp,E. (2008). Drug-efficacy depends on the inhibitor type and the target position in a metabolic network--a systematic study. J. Theor. Biol. *252*, 442-455.

Gidalevitz,T., Ben-Zvi,A., Ho,K.H., Brignull,H.R., and Morimoto,R.I. (2006). Progressive disruption of cellular protein folding in models of polyglutamine diseases. Science *311*, 1471-1474.

Girvan,M. and Newman,M.E. (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. U. S. A *99*, 7821-7826.

Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M., and Barabasi,A.L. (2007). The human disease network. Proc. Natl. Acad. Sci. U. S. A *104*, 8685-8690.

Goldberg,A.L. (2003). Protein degradation and protection against misfolded or damaged proteins. Nature *426*, 895-899.

Guarente,L. (1993). Synthetic enhancement in gene interaction: a genetic tool come of age. Trends Genet. *9*, 362-366.

Guisbert,E., Yura,T., Rhodius,V.A., and Gross,C.A. (2008). Convergence of molecular, modeling, and systems approaches for an understanding of the Escherichia coli heat shock response. Microbiol. Mol. Biol. Rev *72*, 545-554.

Gunther,S., Kuhn,M., Dunkel,M., Campillos,M., Senger,C., Petsalaki,E., Ahmed,J., Urdiales,E.G., Gewiess,A., Jensen,L.J., Schneider,R., Skoblo,R., Russell,R.B., Bourne,P.E., Bork,P., and Preissner,R. (2008). SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res. *36*, D919-D922.

Han,J.D. (2008). Understanding biological functions through molecular networks. Cell Res. *18*, 224-237.

Han,J.D., Bertin,N., Hao,T., Goldberg,D.S., Berriz,G.F., Zhang,L.V., Dupuy,D., Walhout,A.J., Cusick,M.E., Roth,F.P., and Vidal,M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature *430*, 88-93.

Hartman,J.L., Garvik,B., and Hartwell,L. (2001). Principles for the buffering of genetic variation. Science *291*, 1001-1004.

Hipkiss,A.R. (2006). Accumulation of altered proteins and ageing: causes and effects. Exp. Gerontol. *41*, 464-473.

Hoeller,D. and Dikic,I. (2009). Targeting the ubiquitin system in cancer therapy. Nature *458*, 438-444.

Holmes,E., Loo,R.L., Stamler,J., Bictash,M., Yap,I.K., Chan,Q., Ebbels,T., De,I.M., Brown,I.J., Veselkov,K.A., Daviglus,M.L., Kesteloot,H., Ueshima,H., Zhao,L., Nicholson,J.K., and Elliott,P. (2008a). Human metabolic phenotype diversity and its association with diet and blood pressure. Nature *453*, 396-400.

Holmes,E., Wilson,I.D., and Nicholson,J.K. (2008b). Metabolic phenotyping in health and disease. Cell *134*, 714-717.

Hopkins,A.L. (2008). Network pharmacology: the next paradigm in drug discovery. Nat Chem. Biol. *4*, 682-690.

Houle,D., Govindaraju,D.R., and Omholt,S. (2010). Phenomics: the next challenge. Nat Rev Genet. *11*, 855-866.

Ioannidis,J.P. (2005). Why most published research findings are false. PLoS. Med. *2*, e124.

Ioannidis,J.P. and Khoury,M.J. (2011). Improving validation practices in “omics” research. Science *334*, 1230-1232.

Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F., and Gerstein,M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449-453.

Jelier,R., ‘t Hoen,P.A., Sterrenburg,E., den Dunnen,J.T., van Ommen,G.J., Kors,J.A., and Mons,B. (2008). Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. BMC. Bioinformatics. *9*, 291.

Jeong,H., Mason,S.P., Barabasi,A.L., and Oltvai,Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41-42.

Jerby,L., Shlomi,T., and Ruppin,E. (2010). Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol. Syst. Biol. *6*, 401.

Ji,H.F., Kong,D.X., Shen,L., Chen,L.L., Ma,B.G., and Zhang,H.Y. (2007). Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. Genome Biol. *8*, R176.

Jonikas,M.C., Collins,S.R., Denic,V., Oh,E., Quan,E.M., Schmid,V., Weibezahn,J., Schwappach,B., Walter,P., Weissman,J.S., and Schuldiner,M. (2009). Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. Science *323*, 1693-1697.

Joyce,A.R. and Palsson,B.O. (2006). The model organism as a system: integrating ‘omics’ data sets. Nat Rev Mol. Cell Biol. *7*, 198-210.

Kaplan, W. and Laing, R. Priority Medicines for Europe and the World. 2004. Geneva, World Health Organization.

Ref Type: Report

Kent-Braun,J.A., Ng,A.V., Doyle,J.W., and Towse,T.F. (2002). Human skeletal muscle responses vary with age and gender during fatigue due to incremental isometric exercise. J. Appl. Physiol *93*, 1813-1823.

Kim,W., Bennett,E.J., Huttlin,E.L., Guo,A., Li,J., Possemato,A., Sowa,M.E., Rad,R., Rush,J., Comb,M.J., Harper,J.W., and Gygi,S.P. (2011). Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol. Cell *44*, 325-340.

Kirkwood,T.B. (2005). Understanding the odd science of aging. Cell *120*, 437-447.

Kirkwood,T.B. and Austad,S.N. (2000). Why do we age? Nature *408*, 233-238.

Kirouac,D.C., Ito,C., Csaszar,E., Roch,A., Yu,M., Sykes,E.A., Bader,G.D., and Zandstra,P.W. (2010). Dynamic interaction networks in a hierarchically organized tissue. Mol. Syst. Biol. *6*, 417.

Kitano,H. (2007). A robustness-based approach to systems-oriented drug design. Nat Rev Drug Discov. *6*, 202-210.

Kostek,M.C. and Delmonico,M.J. (2011). Age-Related Changes in Adult Muscle Morphology. Curr. Aging Sci.

Krishnamurthy,J., Torrice,C., Ramsey,M.R., Kovalev,G.I., Al-Regaiey,K., Su,L., and Sharpless,N.E. (2004). Ink4a/Arf expression is a biomarker of aging. J. Clin. Invest *114*, 1299-1307.

Kuhn,U., Gundel,M., Knoth,A., Kerwitz,Y., Rudel,S., and Wahle,E. (2009). Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J. Biol. Chem. *284*, 22803-22814.

Lage,K., Mollgard,K., Greenway,S., Wakimoto,H., Gorham,J.M., Workman,C.T., Bendsen,E., Hansen,N.T., Rigina,O., Roque,F.S., Wiese,C., Christoffels,V.M., Roberts,A.E., Smoot,L.B., Pu,W.T., Donahoe,P.K., Tommerup,N., Brunak,S., Seidman,C.E., Seidman,J.G., and Larsen,L.A. (2010). Dissecting spatio-temporal protein networks driving human heart development and related disorders. Mol. Syst. Biol. *6*, 381.

Lee,C.K., Klopp,R.G., Weindruch,R., and Prolla,T.A. (1999). Gene expression profile of aging and its retardation by caloric restriction. Science *285*, 1390-1393.

Lemay,J.F., D'Amours,A., Lemieux,C., Lackner,D.H., St-Sauveur,V.G., Bahler,J., and Bachand,F. (2010). The nuclear poly(A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. Mol. Cell *37*, 34-45.

Lemieux,C. and Bachand,F. (2009). Cotranscriptional recruitment of the nuclear poly(A)-binding protein Pab2 to nascent transcripts and association with translating mRNPs. Nucleic Acids Res. *37*, 3418-3430.

Lexell,J., Taylor,C.C., and Sjostrom,M. (1988). What is the cause of the ageing atrophy? Total number, size and proportion of different fiber types studied in whole vastus lateralis muscle from 15- to 83-year-old men. J. Neurol. Sci. *84*, 275-294.

Lindle,R.S., Metter,E.J., Lynch,N.A., Fleg,J.L., Fozard,J.L., Tobin,J., Roy,T.A., and Hurley,B.F. (1997). Age and gender comparisons of muscle strength in 654 women and men aged 20-93 yr. J. Appl. Physiol *83*, 1581-1587.

Linoli,G., Tomelleri,G., and Ghezzi,M. (1991). Oculopharyngeal muscular dystrophy. Description of a case with involvement of the central nervous system]. Pathologica *83*, 325-334.

Lipkowitz,S. and Weissman,A.M. (2011). RINGs of good and evil: RING finger ubiquitin ligases at the crossroads of tumour suppression and oncogenesis. Nat Rev Cancer *11*, 629-643.

Liu,Z., Miers,W.R., Wei,L., and Barrett,E.J. (2000). The ubiquitin-proteasome proteolytic pathway in heart vs skeletal muscle: effects of acute diabetes. Biochem. Biophys. Res. Commun. *276*, 1255-1260.

Lu,T., Pan,Y., Kao,S.Y., Li,C., Kohane,I., Chan,J., and Yankner,B.A. (2004). Gene regulation and DNA damage in the ageing human brain. Nature *429*, 883-891.

Lu,Y., Huggins,P., and Bar-Joseph,Z. (2009). Cross species analysis of microarray expression data. Bioinformatics. *25*, 1476-1483.

Luscombe,N.M., Babu,M.M., Yu,H., Snyder,M., Teichmann,S.A., and Gerstein,M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308-312.

Martin,G.M. (2009). Epigenetic gambling and epigenetic drift as an antagonistic pleiotropic mechanism of aging. Aging Cell *8*, 761-764.

Medici,M., Pizzarossa,C., Skuk,D., Yorio,D., Emmanuelli,G., and Mesa,R. (1997). Oculopharyngeal muscular dystrophy in Uruguay. Neuromuscul. Disord. *7 Suppl 1*, S50-S52.

Meusser,B., Hirsch,C., Jarosch,E., and Sommer,T. (2005). ERAD: the long road to destruction. Nat Cell Biol. *7*, 766-772.

Millefiorini,M. and Filippini,C. (1967). Oculopharyngeal muscular dystrophy. Riv. Neurol. *37*, 327-337.

Mizoi,Y., Yamamoto,T., Minami,N., Ohkuma,A., Nonaka,I., Nishino,I., Tamura,N., Amano,T., and Araki,N. (2011). Oculopharyngeal muscular dystrophy associated with dementia. Intern. Med. *50*, 2409-2412.

Moore,M.J. and Silver,P.A. (2008). Global analysis of mRNA splicing. RNA. *14*, 197-203.

Morimoto,R.I. (2008). Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging. Genes Dev. *22*, 1427-1438.

Munch,C. and Bertolotti,A. (2010). Exposure of hydrophobic surfaces initiates aggregation of diverse ALS-causing superoxide dismutase-1 mutants. J. Mol. Biol. *399*, 512-525.

Nalepa,G., Rolfe,M., and Harper,J.W. (2006). Drug discovery in the ubiquitin-proteasome system. Nat Rev Drug Discov. *5*, 596-613.

Nicholson,J.K. and Wilson,I.D. (2003). Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. Nat Rev Drug Discov. *2*, 668-676.

Oberdoerffer,P. and Sinclair,D.A. (2007). The role of nuclear architecture in genomic instability and ageing. Nat Rev Mol. Cell Biol. *8*, 692-702.

Ochs,M.F. (2010). Knowledge-based data analysis comes of age. Brief. Bioinform. *11*, 30-39.

Oliva,A., Rosebrock,A., Ferrezuelo,F., Pyne,S., Chen,H., Skiena,S., Futcher,B., and Leatherwood,J. (2005). The cell cycle-regulated genes of Schizosaccharomyces pombe. PLoS. Biol. *3*, e225.

Olzmann,J.A., Li,L., Chudaev,M.V., Chen,J., Perez,F.A., Palmiter,R.D., and Chin,L.S. (2007). Parkin-mediated K63-linked polyubiquitination targets misfolded DJ-1 to aggresomes via binding to HDAC6. J. Cell Biol. *178*, 1025-1038.

Palla,G., Derenyi,I., Farkas,I., and Vicsek,T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature *435*, 814-818.

Pan,Q., Shai,O., Lee,L.J., Frey,B.J., and Blencowe,B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. *40*, 1413-1415.

Pe'er,D., Regev,A., and Tanay,A. (2002). Minreg: inferring an active regulator set. Bioinformatics. *18 Suppl 1*, S258-S267.

Perie,S., Mamchaoui,K., Mouly,V., Blot,S., Bouazza,B., Thornell,L.E., St Guily,J.L., and Butler-Browne,G. (2006). Premature proliferative arrest of cricopharyngeal myoblasts in oculo-pharyngeal muscular dystrophy: Therapeutic perspectives of autologous myoblast transplantation. Neuromuscul Disord *16*, 770-781.

Pogribny,I.P., Ross,S.A., Tryndyak,V.P., Pogribna,M., Poirier,L.A., and Karpinets,T.V. (2006). Histone H3 lysine 9 and H4 lysine 20 trimethylation and the expression of Suv4-20h2 and Suv-39h1 histone methyltransferases in hepatocarcinogenesis induced by methyl deficiency in rats. Carcinogenesis *27*, 1180-1186.

Potapov,A.P., Goemann,B., and Wingender,E. (2008). The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks. BMC. Bioinformatics. *9*, 227.

Powers,E.T., Morimoto,R.I., Dillin,A., Kelly,J.W., and Balch,W.E. (2009). Biological and chemical approaches to diseases of proteostasis deficiency. Annu. Rev Biochem. *78*, 959-991.

Rajan,R.S., Illing,M.E., Bence,N.F., and Kopito,R.R. (2001). Specificity in intracellular protein aggregation and inclusion body formation. Proc. Natl. Acad. Sci. U. S. A *98*, 13060-13065.

Rakyan,V.K., Down,T.A., Maslau,S., Andrew,T., Yang,T.P., Beyan,H., Whittaker,P., McCann,O.T., Finer,S., Valdes,A.M., Leslie,R.D., Deloukas,P., and Spector,T.D. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. *20*, 434-439.

Rando,T.A. and Chang,H.Y. (2012). Aging, rejuvenation, and epigenetic reprogramming: resetting the aging clock. Cell *148*, 46-57.

Raz,V., Routledge,S., Venema,A., Buijze,H., van der Wal,E., Anvar,S.Y., Straasheijm,K.R., Klooster,R., Antoniou,M., and van der Maarel,S.M. (2011). Modeling Oculopharyngeal Muscular Dystrophy in Myotube Cultures Reveals Reduced Accumulation of Soluble Mutant PABPN1 Protein. Am. J. Pathol.

Reshef,D.N., Reshef,Y.A., Finucane,H.K., Grossman,S.R., McVean,G., Turnbaugh,P.J., Lander,E.S., Mitzenmacher,M., and Sabeti,P.C. (2011). Detecting novel associations in large data sets. Science *334*, 1518-1524.

Reverter,A., Ingham,A., and Dalrymple,B.P. (2008). Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. BioData. Min *1*, 8.

Ron,D. and Walter,P. (2007). Signal integration in the endoplasmic reticulum unfolded protein response. Nat Rev Mol. Cell Biol. *8*, 519-529.

Ronn,T., Poulsen,P., Hansson,O., Holmkvist,J., Almgren,P., Nilsson,P., Tuomi,T., Isomaa,B., Groop,L., Vaag,A., and Ling,C. (2008). Age influences DNA methylation and gene expression of COX7A1 in human skeletal muscle. Diabetologia *51*, 1159-1168.

Roth,S.M., Ferrell,R.E., Peters,D.G., Metter,E.J., Hurley,B.F., and Rogers,M.A. (2002). Influence of age, sex, and strength training on human muscle gene expression determined by microarray. Physiol Genomics *10*, 181-190.

Saeed,R. and Deane,C.M. (2006). Protein protein interactions, evolutionary rate, abundance and age. BMC. Bioinformatics. *7*, 128.

Sahin,E. and Depinho,R.A. (2010). Linking functional decline of telomeres, mitochondria and stem cells during ageing. Nature *464*, 520-528.

Sandri,M. (2008). Signaling in muscle atrophy and hypertrophy. Physiology. (Bethesda.) *23*, 160-170.

Sarkar,A.K., Biswas,S.K., Ghosh,A.K., Mitra,P., Ghosh,S.K., and Mathew,J. (1995). Oculopharyngeal muscular dystrophy. Indian J. Pediatr. *62*, 496-498.

Schieppati,A., Henter,J.I., Daina,E., and Aperia,A. (2008). Why rare diseases are an important medical and social issue. Lancet *371*, 2039-2041.

Schilling,C.H., Edwards,J.S., and Palsson,B.O. (1999). Toward metabolic phenomics: analysis of genomic data using flux balances. Biotechnol. Prog. *15*, 288-295.

Schuldiner,M., Metz,J., Schmid,V., Denic,V., Rakwalska,M., Schmitt,H.D., Schwappach,B., and Weissman,J.S. (2008). The GET complex mediates insertion of tail-anchored proteins into the ER membrane. Cell *134*, 634-645.

Searls,D.B. (2005). Data integration: challenges for drug discovery. Nat Rev Drug Discov. *4*, 45-58.

Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D., and Friedman,N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. *34*, 166-176.

Segre,D., Deluna,A., Church,G.M., and Kishony,R. (2005). Modular epistasis in yeast metabolism. Nat Genet. *37*, 77-83.

Sexton,T., Schober,H., Fraser,P., and Gasser,S.M. (2007). Gene regulation through nuclear organization. Nat Struct. Mol. Biol. *14*, 1049-1055.

Soule,M. (1967). Phenetics of Natural Populations I. Phenetic Relationships of Insular Populations of the Side-Blotched Lizard. Evolution *21*, 584-591.

Spector,D.L. and Gasser,S.M. (2003). A molecular dissection of nuclear function. Conference on the dynamic nucleus: questions and implications. EMBO Rep. *4*, 18-23.

Spiro,Z., Kovacs,I.A., and Csermely,P. (2008). Drug-therapy networks and the prediction of novel drug targets. J. Biol. *7*, 20.

Sprinzak,E., Sattath,S., and Margalit,H. (2003). How reliable are experimental protein-protein interaction data? J. Mol. Biol. *327*, 919-923.

St Onge,R.P., Mani,R., Oh,J., Proctor,M., Fung,E., Davis,R.W., Nislow,C., Roth,F.P., and Giaever,G. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nat Genet. *39*, 199-206.

Steele,E., Tucker,A., 't Hoen,P.A., and Schuemie,M.J. (2009). Literature-based priors for gene regulatory networks. Bioinformatics. *25*, 1768-1774.

Stolk,P., Willemen,M.J., and Leufkens,H.G. (2006). Rare essentials: drugs for rare diseases as essential medicines. Bull. World Health Organ *84*, 745-751.

Taillandier,D., Combaret,L., Pouch,M.N., Samuels,S.E., Bechet,D., and Attaix,D. (2004). The role of ubiquitin-proteasome-dependent proteolysis in the remodelling of skeletal muscle. Proc. Nutr. Soc. *63*, 357-361.

Tanay,A., Sharan,R., Kupiec,M., and Shamir,R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc. Natl. Acad. Sci. U. S. A *101*, 2981-2986.

Tazi,J., Bakkour,N., and Stamm,S. (2009). Alternative splicing and disease. Biochim. Biophys. Acta *1792*, 14-26.

Teschendorff,A.E., Menon,U., Gentry-Maharaj,A., Ramus,S.J., Weisenberger,D.J., Shen,H., Campan,M., Noushmehr,H., Bell,C.G., Maxwell,A.P., Savage,D.A., Mueller-Holzner,E., Marth,C., Kocjan,G., Gayther,S.A., Jones,A., Beck,S., Wagner,W., Laird,P.W., Jacobs,I.J., and Widschwendter,M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. *20*, 440-446.

Tibshirani,R. (1996). Regression Shrinkage and Selection via the LASSO. Journal of the Royal Statistical Society *58*, 267-288.

Tohgi,H., Utsugisawa,K., Nagane,Y., Yoshimura,M., Genda,Y., and Ukitsu,M. (1999). Reduction with age in methylcytosine in the promoter region -224 approximately -101 of the amyloid precursor protein gene in autopsy human cortex. Brain Res. Mol. Brain Res. *70*, 288-292.

Tong,A.H., Lesage,G., Bader,G.D., Ding,H., Xu,H., Xin,X., Young,J., Berriz,G.F., Brost,R.L., Chang,M., Chen,Y., Cheng,X., Chua,G., Friesen,H., Goldberg,D.S., Haynes,J., Humphries,C., He,G., Hussein,S., Ke,L., Krogan,N., Li,Z., Levinson,J.N., Lu,H., Menard,P., Munyana,C., Parsons,A.B., Ryan,O., Tonikian,R., Roberts,T., Sdicu,A.M., Shapiro,J., Sheikh,B., Suter,B., Wong,S.L., Zhang,L.V., Zhu,H., Burd,C.G., Munro,S., Sander,C., Rine,J., Greenblatt,J., Peter,M., Bretscher,A., Bell,G., Roth,F.P., Brown,G.W., Andrews,B., Bussey,H., and Boone,C. (2004). Global mapping of the yeast genetic interaction network. Science *303*, 808-813.

Trinkle-Mulcahy,L. and Lamond,A.I. (2007). Toward a high-resolution view of nuclear dynamics. Science *318*, 1402-1407.

Trollet,C., Anvar,S.Y., Venema,A., Hargreaves,I.P., Foster,K., Vignaud,A., Ferry,A., Negroni,E., Hourde,C., Baraibar,M.A., 't Hoen,P.A., Davies,J.E., Rubinsztein,D.C., Heales,S.J., Mouly,V., van der Maarel,S.M., Butler-Browne,G., Raz,V., and Dickson,G. (2010). Molecular and phenotypic characterization of a mouse model of oculopharyngeal muscular dystrophy reveals severe muscular atrophy restricted to fast glycolytic fibres. Hum. Mol. Genet.

Tryndyak,V.P., Kovalchuk,O., and Pogribny,I.P. (2006). Loss of DNA methylation and histone H4 lysine 20 trimethylation in human breast cancer cells is associated with aberrant expression of DNA methyltransferase 1, Suv4-20h2 histone methyltransferase and methyl-binding proteins. Cancer Biol. Ther. *5*, 65-70.

Tyedmers,J., Mogk,A., and Bukau,B. (2010). Cellular strategies for controlling protein aggregation. Nat Rev Mol. Cell Biol. *11*, 777-788.

Venkatesan,K., Rual,J.F., Vazquez,A., Stelzl,U., Lemmens,I., Hirozane-Kishikawa,T., Hao,T., Zenkner,M., Xin,X., Goh,K.I., Yildirim,M.A., Simonis,N., Heinzmann,K., Gebreab,F., Sahalie,J.M., Cevik,S., Simon,C., de Smet,A.S., Dann,E., Smolyar,A., Vinayagam,A., Yu,H., Szeto,D., Borick,H., Dricot,A., Klitgord,N., Murray,R.R., Lin,C., Lalowski,M., Timm,J., Rau,K., Boone,C., Braun,P., Cusick,M.E., Roth,F.P., Hill,D.E., Tavernier,J., Wanker,E.E., Barabasi,A.L., and Vidal,M. (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Vignaud,A., Ferry,A., Huguet,A., Baraibar,M., Trollet,C., Hyzewicz,J., Butler-Browne,G., Puymirat,J., Gourdon,G., and Furling,D. (2010). Progressive skeletal muscle weakness in transgenic mice expressing CTG expansions is associated with the activation of the ubiquitin-proteasome pathway. Neuromuscul. Disord. *20*, 319-325.

Wahle,E. (1991). A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. Cell *66*, 759-768.

Wahle,E. (1995). Poly(A) tail length control is caused by termination of processive synthesis. J. Biol. Chem. *270*, 2800-2808.

Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P., and Burge,C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang,G.S. and Cooper,T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet. *8*, 749-761.

Wang,J., Farr,G.W., Zeiss,C.J., Rodriguez-Gil,D.J., Wilson,J.H., Furtak,K., Rutkowski,D.T., Kaufman,R.J., Ruse,C.I., Yates,J.R., III, Perrin,S., Feany,M.B., and Horwich,A.L. (2009). Progressive aggregation despite chaperone associations of a mutant SOD1-YFP in transgenic mice that develop ALS. Proc. Natl. Acad. Sci. U. S. A *106*, 1392-1397.

Wunderlich,Z. and Mirny,L.A. (2006). Using the topology of metabolic networks to predict viability of mutant strains. Biophys. J. *91*, 2304-2311.

Xu,G. and Jaffrey,S.R. (2011). The new landscape of protein ubiquitination. Nat Biotechnol. *29*, 1098-1100.

Yildirim,M.A., Goh,K.I., Cusick,M.E., Barabasi,A.L., and Vidal,M. (2007). Drug-target network. Nat Biotechnol. *25*, 1119-1126.

Yu,H., Kim,P.M., Sprecher,E., Trifonov,V., and Gerstein,M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS. Comput. Biol. *3*, e59.

Zahn,J.M., Sonu,R., Vogel,H., Crane,E., Mazan-Mamczarz,K., Rabkin,R., Davis,R.W., Becker,K.G., Owen,A.B., and Kim,S.K. (2006). Transcriptional profiling of aging in human muscle reveals a common aging signature. PLoS. Genet. *2*, e115.

Zhang,C., Frias,M.A., Mele,A., Ruggiu,M., Eom,T., Marney,C.B., Wang,H., Licatalosi,D.D., Fak,J.J., and Darnell,R.B. (2010). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Science *329*, 439-443.

Zhang,L.V., King,O.D., Wong,S.L., Goldberg,D.S., Tong,A.H., Lesage,G., Andrews,B., Bussey,H., Boone,C., and Roth,F.P. (2005). Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J. Biol. *4*, 6.

Zhou,X.J. and Gibson,G. (2004). Cross-species comparison of genome-wide expression patterns. Genome Biol. *5*, 232.

summary

Long after the discovery of DNA by Friedrich Miescher in 1869 and recognition of its central role as holder of genetic information, we still know little about how DNA is determining the development and function of all cells and organisms. . Within the double helix structure of the DNA, discovered by James D. Watson and Francis Crick in 1953, it is the four nucleotides (Adenine, Thymine, Cytosine, and Guanine) that carry the genetic information. It is astonishing that the level of complexity and diversity among living organisms, or between objects of the same species, arise from differing sequences of the four nucleotides The interpretation of this highly diverse and evolving genomic information becomes more crucial as we intend to better understand the molecular mechanisms underlying human disorders. By the early 1970s, the growing concept and advancements by Fred Sanger in sequencing the first genome (Φ-X174 bacteriophage) established the cornerstones of innovations in 21st century systems biology. The field of systems biology is aimed to model how individual elements of the cell interact in a concerted fashion to bring forth highly dynamic biological organisation and behaviours in versatile environments. To many scientists, the rise of systems biology goes back to the beginning of the last decade, owing to the advent of high-throughput technologies in production of vast amount of genomic, transcriptomic, and proteomic data. As genetics is aimed to answer the question of 'what', systems biology is aimed to construct models that are designed to go one step beyond by tackling the question of 'how'.

Complexity is perhaps the most common adjective used to describe biology and its related computational models. In every cell, biological functions are mediated through complex networks of interactions between metabolites, proteins, and DNA. On the basis that cells are evolved to survive and not for scientists to understand, the stochastic nature of biological data requires special efforts for combined and interdisciplinary investigations. Looking for common patterns that underlie the diversity, development, and inner cell dynamics of organisms can lead to uncover the most prominent and shared functional features. Likewise, in the field of computational biology, inspirations from biological systems led to development of novel algorithms for knowledge discovery. Early works on data-mining and machine learning in the 1960s, for instance, evolved around the idea on the activity of neurons in the brain to give rise to a class of powerful algorithms known as neural networks. Genetic algorithm is another example where inspirations from common operations in DNA sequence evolution led to the development of one of the most used optimisation techniques in the field of systems biology. In realisation of complexity and variety of living organisms and biological processes, the 21st century systems biology has ever more embraced the idea of interdisciplinary and combined efforts for knowledge discovery.

In this thesis, I have explored novel strategies that can bridge the gap between multi-layers of biomedical data to provide a strong vision on how molecular networks are structured, under various conditions, to attain particular functional behaviours underlying human diseases. Extensive use of computational and integration approaches can provide comprehensive and more accurate mechanistic insights on the disease pathogenesis. In doing so, we have focused on attempting to unravel molecular mechanisms that are involved in the aetiology of oculopharyngeal muscular dystrophy (OPMD). OPMD is an autosomal dominant and late-onset disorder which usually manifest in midlife, after the age of 40. The main symptoms are slowly progressive *ptosis* (drooping of eyelid), *dysphagia* (difficulty swallowing), and weakness of proximal limb muscles. In most patients, life expectancy is not reduced. However, the quality of life is greatly affected as *ptosis* can cause visual limitations, *dysphagia* may lead to aspiration pneumonia and weight loss, and patients with proximal limb weakness can eventually be wheelchair bound. Cases of OPMD are reported in over 30 countries and the prevalence is estimated to be of 1 in 100,000 worldwide. The genetic cause of this disease is the expansion mutation in the Poly(A) Binding Protein Nuclear 1 (PABPN1) protein. The underlying molecular mechanisms by which the mutated PABPN1 causes tissue-specific and progressive muscle weakness are not fully understood. In chapter **one** and **two**, we have shown that attentive modelling and optimization of integration strategy can serve as a powerful system for knowledge discovery. Combinatory survey of differing expression patterns and collective transcriptional behaviours into structured communities led to the discovery of the ubiquitin-proteasome system as the most prominently involved molecular pathway in OPMD patients and model systems. In addition, our data indicated that age-dependent and progressive decline of PABPN1 expression result in progressive deregulation of muscle contractile genes, induction of cellular senescence, and decline of cell growth and fusion (chapter **three**). Since PABPN1 regulates mRNA stability it is expected that decline in functional PABPN1 would have a broad effect on cellular functions. Our data suggest a progressive response of muscle cell function to the level of PABPN1 in a spatial-temporal manner, highlighting PABPN1 role as a regulator of muscle ageing. Understanding the underlying causes of OPMD is a key step toward enabling earlier and more precise diagnosis, prognosis, therapeutic interventions, drug discovery and potential prevention.

Functional interdependencies and the modular nature of the molecular components of the cell necessitate the study of biological networks in human diseases. Moreover, integration of data and genomic information from human and various model systems can provide a better indication of common molecular mechanisms that underlie a given phenotype. Therefore, I provided a framework in which a model-driven construction of disease networks on modules of functionally related genes can be translated across species to identify the most essential regulatory relationships (chapter **four** and **five**). This is where bridging the multi-layers of biomedical data can transform the field of biomedical network inference and analysis. The adoption of methods that deal with evolutionary dynamics of these networks, in a spatial-temporal manner, can act as a cornerstone for robust integration of pharmaceutical data and chemical interactions. This combinatory strategy provides a valuable framework for drug discovery and personalized therapeutic interventions as our understanding of biological networks and phenotypes plays an essential role in improving efficacy of a drug and inhibiting its off-target toxicity. Improvements in systems biology methodologies will bring us ever closer to the central question of 'how' and beyond.

samenvatting

Lang na de ontdekking van DNA door Friedrich Miesscher in 1869 en de herkenning van DNA als de bron van genetische informatie, weten we nog maar weinig over de manier waarop DNA de ontwikkeling en functie van cellen en organismen bepaalt. Binnen de dubbele helix structuur van het DNA, ontdekt door James D. Watson and Francis Crick in 1953, zijn het de vier nucleotiden (Adenine, Thymine, Cytosine en Guanine) die de genetische informatie dragen. Het is verbazingwekkend dat verschillende opeenvolging van de vier nucleotiden ten grondslag ligt aan de hoge complexiteit and diversiteit tussen organismen, of zelfs tussen individuen van dezelfde soort.bij elkaar gehouden door glucose- en fosfaat-groepen. De interpretatie van de verscheidenheid in genomische informatie is cruciaal wanneer we de moleculaire mechanismen onderliggend aan ziektes proberen te begrijpen. Aan het begin van de jaren 70, legden het concept en de vorderingen van Fred Sanger bij het sequensen van het eerste genoom (Φ-X174 bacteriophage) de grondslag voor de innovaties in de systeem biologie van de 21e eeuw. Het doel van systeem biologie is om een model op te zetten dat laat zien hoe individuele elementen van de cel met elkaar samenwerken in een dynamische biologische organisatie. Volgens veel onderzoekers is het opkomen van systeem biologie tijdens het laatse decennium te danken aan de opkomst van de high-throughput technologie, die in staat is een grote hoeveelheid genomische, transcriptomische en proteomische data te produceren. Terwijl de genetica is gericht op het beantwoorden van de vraag “wat?”, richt de systeem biologie zich op het construeren van modellen die ontworpen zijn om een stap verder te gaan, en de vraag “hoe?” proberen te beantwoorden.

De term complexiteit wordt veelal gebruikt om biologie en de gerelateerde computer modellen te beschrijven. In elke cell worden biologisch functies gemedieerd door complexe netwerken van interacties tussen metabolieten, eiwitten en DNA. Ervan uitgaande dat cellen zijn geëvolueerd om te overleven en niet om begrepen te kunnen worden door onderzoekers, vraagt de stochastische natuur van biologische data om een combinatie van specialistische inzet en interdicliplinair onderzoek. Het zoeken naar patronen die ten grondslag liggen aan de diversiteit, ontwikkeling en intracellulaire dynamiek van organismen kan leiden tot de ontdekking van de meest vooraanstaande en gedeelde functionele eigenschappen. In het veld van de computationele biologie heeft inspiratie uit biologische systemen geleid tot de ontwikkeling van nieuwe algoritmen voor kennisvergaring (. Zo heeft werk aan ‘data-mining’ en ‘machine learning’ tijdens de jaren 60, gebasseerd op het idee van de activiteit van neuronen in de hersenen, geleid tot de ontwikkeling van krachtige algoritmen, genaamd ‘neurale netwerken’. Genetisch algoritmen zijn andere voorbeelden, waarbij inspiratie uit gemeenschappelijke functies in DNA sequentie evolutie hebben geleid to de ontwikkeling van een van de meest gebruikte optimalisatie technieken in het veld van systeem biologie. In de herkenning van de complexiteit en verscheidenheid van levende organismen en biologische processen, heeft de systeem biologie van de 21e eeuw het idee van interdiciplinair en gecombineerd onderzoek voor kennisvergaring steeds meer omarmd.

In dit proefschrift, heb ik nieuwe strategieën onderzocht, die een brug slaan tussen de diverse lagen van biomedische data. Hierbij was het doel inzicht te verschaffen in de structuur van moleculaire netwerken onder verschillende condities, om zo de functionele eigenschappen die ten grondslag liggen aan menselijke aandoeningen te achterhalen. Uitgebreid gebruik van computationele en integratieve benaderingen kunnen nauwkeuriger mechanistische inzichten geven in de pathogenese van menselijke aandoeningen. Op deze manier, heb ik mij gericht op de ontrafeling van moleculaire mechanismen die betrokken zijn bij de etiologie van oculopharyngeale spier dystrofie (OPMD). OPMD is een autosomaal dominante aandoening die tot uiting komt na het 40e levensjaar. De belangrijkste symptomen van OPMD zijn progressieve ptosis (hangende oogleden), dysfagie (moeite met slikken) en spierzwakte van proximale ledematen. Hoewel bij de meeste patienten de levensverwachting normaal is, wordt de kwaliteit van leven ernstig beinvloed, aangezien ptosis visuele beperkingen kan veroorzaken, dysfagie kan leiden tot aspiratiepneumonie en gewichtsverlies en patiënten met spierzwakte van de proximale ledematen uiteindelijk afhanklijk kunnen worden van een rolstoel. Meer dan 30 landen hebben gevallen van OPMD gemeld en de prevalentie wordt geschat op 1 op de 100.000. De genetische oorzaak van deze aandoening is een mutatie in het Poly(A) Binding Protein Nuclear 1 (PABPN1) eiwit. Het onderliggende moleculaire mechanisme waarmee het gemuteerde PABPN1 eiwit progressieve spierzwakte veroorzaakt is niet duidelijk. In hoofdstuk **één** en **twee** hebben we laten zien dat aandachtig modelleren en optimaliseren van integratie strategieën nieuwe mechanistische kennis kan opleveren. Gecombineerd onderzoek naar de veranderlijke expressie patronen heeft geleid tot de ontdekking van het ubiquitine-proteasoom systeem, als het meest prominent betrokken moleculaire systeem in OPMD patiënten en model systemen. Onze data suggereren dat leeftijd-afhankelijke en progressieve afname van PABPN1 expressie resulteert in deregulatie van spier contractie genen, inductie van cel veroudering en afname van cel groei en fusie (hoofdstuk **drie**). Omdat PABPN1 mRNA stabiliteit reguleert, is het te verwachten dat een afname van functioneel PABPN1 eiwit een breed effect heeft op cellulaire functies. Als eerste in dit veld, laat onze data het progressieve effect van het niveau van PABPN1 op de functie van spier cellen zien, en benadrukt het de rol van PABPN1 als een regulator van spier veroudering. Het begrijpen van de onderliggende oorzaken van OPMD is een belangrijke stap richting eerdere en meer precieze diagnose, prognose, behandeling, medicijn ontwikkeling en het eventueel voorkomen van de ziekte.

De onderlinge functionele afhankelijkheid en de modulaire aard van de moleculaire componenten van de cel maken de studie van biologische netwerken in menselijke aandoeningen noodzakelijk. Daarnaast geeft de integratie van data en genomische informatie van de mens en verscheidene model systemen een betere indicatie van gemeenschappelijke moleculaire mechanimsen die ten grondslag liggen aan een bepaald fenotype. Daarom heb ik een kader verschaft waarin een model gedreven constructie van ziekte netwerken op basis van modules van functioneel gerelateerde genen zijn getransleerd over verschillende diersoorten om de meest essentiële relaties te identificeren (Hoofdstuk **vier** en **vijf**). De brug tussen de diverse lagen van biomedische data kan zo het veld van biomedische netwerk transformeren. Methoden die de evolutionaire dynamiek van deze netwerken modelleren, in een ruimte- en tijdsafhankelijke manier, vormen een basis voor de robuuste integratie van farmaceutische data en chemische interacties. Deze gecombineerde strategie kan ook waardevol zijn voor medicijn ontwikkeling en geindividualiseerde behandeling. Besef van biologische netwerken en fenotypes zal een essentiële rol spelen in het verbeteren van de werkzaamheid van medicijnen en het verminderen van bijwerkingen. Nieuwe methoden uit de systeem biologie zullen ons helpen bij het beantwoorden van de centrale vraag “hoe?” en verder.

abbreviations

1	**1PB**	*one parent Bayesian network*
A	**A**	*adenine*
	Ala	*alanine*
B	**BIC**	*Bayes information criterion*
	BNC	*Bayesian network classifier*
C	**C**	*cytosine*
	CDF	*cumulative distribution function*
	CS	*citrate synthase*
	CSA	*cross-sectional area*
	CV	*cross-validation*
D	**D.E.**	*differentially expressed*
	DAG	*directed acyclic graph*
	DAVID	*database for annotation, visualization, and integrated discovery*
	DNA	*deoxyribonucleic acid*
	DUB	*deubiquitinating enzyme*
E	**E1**	*ubiquitin-activating enzyme*
	E2	*ubiquitin-conjugating enzyme*
	E3	*ubiquitin ligase*
	EDL	*extensor digitorum longus muscle*
	EF	*embryonic fibroblast*
	expPABPN1	*expanded Poly(A) Binding Protein Nuclear 1*
F	**FDR**	*false discovery rate*

	FVB	*friend virus B inbred*
G	**G**	*guanine*
	GO	*gene ontology*
	GT	*global test*
H	**HF**	*heterochromatic foci*
I	**IND**	*independent*
	INI	*intranuclear inclusion*
K	**KEGG**	*Kyoto Encyclopedia of genes and genomes*
	KS	*Kolmogorov-Smirnov test*
L	**LAS**	*literature-aided association study*
M	**MCMC**	*Markov chain Monte Carlo*
	MDIC	*multiple datasets with increasing complexity*
	MIC	*maximal information coefficient*
	mRNA	*messenger ribonucleic acid*
	MyHC	*myosin heavy chain*
N	**NBC**	*naïve Bayes classifier*
	ND	*non-deregulated*
	NPB	*unlimited Bayesian network*
	NS	*not significant*
	NT	*non-transduced*
O	**OPMD**	*oculopharyngeal muscular dystrophy*
P	**PABPN1**	*poly(A) binding protein nuclear 1*
	PCA	*principal component analysis*
	PPI	*protein-protein interaction*
R	**RIN**	*RNA integration number*
	RNA-Seq	*RNA sequencing*
	RT qPCR	*reverse transcription polymerase chain reaction*
S	**SNB**	*selective naïve Bayes*
	SOL	*soleus muscle*
	SSE	*sum squared error*
T	**T**	*thymine*

	TA	*tibialis anterior muscle*
	TAN	*tree augmented network*
U	**Ub**	*ubiquitin*
	UPS	*ubiquitin-proteasome system*
	UTR	*un-translated region*
W	**WT**	*wild-type*
Y	**YFP**	*yellow fluorescent protein*

curriculum vitae

Seyed Yahya Anvar was born on October 11, 1980, in Tehran, Iran. He attended *Roozbeh High School*, specialised in Mathematics and Physics, in Tehran and graduated in the summer of 1998. He then was admitted to the Industrial Engineering programme at *Azad University*, Tehran. During the period from 1998 to 2001, he worked as a member of design and development team at Book City Co., project monitoring and consultancy at Negah Multimedia Co., and system manager and web administrator at Book City online stores. From fall 2002 until July 2006, he did a Bachelors of Information Technology at the *Eastern Mediterranean University* located in Famagusta (Gazimağusa), Cyprus, and graduated with high-honour. During this time he received seven high-honour scholarships and one honour scholarship from the Rector.

In September 2006 he began a Masters of Bioinformatics at the *Brunel University* in London, UK, and received his Masters with distinction in 2007. The work presented as part of his Masters dissertation, entitled *Incremental Bayesian Network Models*, led to his participation in collaboration between dr. Allan Tucker at the *Brunel University* and dr. Peter-Bram 't Hoen at the *Leiden University Medical Center*. He then began to develop new methods to apply Bayesian networks for modelling multiple transcriptome datasets related to *myogenesis*.

Starting November 2008 he began a PhD at *Leiden University*, situated at the Center for Human and Clinical Genetics at the *Leiden University Medical Center* under the supervision of Prof. dr. Silvère van der Maarel and co-supervision of dr. Peter-Bram 't Hoen, dr. Vered Raz, and dr. Allan Tucker. The PhD focused on the fascinating nature and value of interdisciplinary studies of human disorders. In this thesis, Yahya and colleagues have shown that engaging in the study of diverse biological systems, in the light of evolutionary mechanisms and shared molecular features (in a rare disease such as oculopharyngeal muscular dystrophy), enabled them to uncover insights on a broad spectrum of conditions and phenomena such as ageing of skeletal muscles and protein aggregation disorders. To do this, he collaborated with centres such as *Brunel University* (dr. Allan Tucker and dr. Veronica Vinciotti), *Radboud University Nijmegen Medical Center* (Prof. dr. Baziel van Engelen), *University of Copenhagen* (dr. John Vissing), and *Royal Holloway University of London* (Prof. dr. George Dickson). During this work, he attended and presented his work at numerous national and international conferences including *Netherlands Bioinformatics Conference (NBIC)*, *European Molecular Biology Organization (EMBO)*, *European Society of Human Genetics (ESHG)*, *American Society of Human Genetics and International Congress of Human Genetics (ASHG/ICHG)*, and *International Conference on Intelligent Systems for Molecular Biology (ISMB)*.

During his time at LUMC, Yahya has reviewed for number of journals including *Bioinformatics*, *BMC Bioinformatics*, *PLoS ONE*, and *BMC Genomics*. He was involved in *Bioinformatics, Computational Biology of Complex diseases and Ageing* (FOS) workshop and MGC course on *Technology Facilities*. He has also helped organising the *Human and Clinical Genetics Party 2009* and *Human and Clinical Genetics Day Out 2010*.

On May 1, 2011 he received a postdoctoral position at Leiden University Medical Center, Center for Human and Clinical Genetics. He now moves on to the intriguing era of the state-of-the-art genome and transcriptome sequencing. He continues with developing new methodologies to model associations and regulatory relationships. These advancements help to better understand the global organization of networks and biological responses in the study of human genetic disorders and other biological conditions. In a slightly different strand, he also has great interests in fine arts, literature, and sports.

publication list

MANUSCRIPT SUBMITTED, 2012

Are ‘identical’ twins identical? Whole genome sequencing of centenarian monozygous twins. K Ye, M Beekman, EW Lameijer, Y Zhang, E van den Akker, J Deelen, JJ Houwing-Duistermaat, D Kremer, SY Anvar, JFJ Laros, D Jones, K Raine, B Blackburne, S Potluri, Q Long, V Guryev, R van der Breggen, R Westendorp, PAC ’t Hoen, JT den Dunnen, GJ van Ommen, G Willemse, DR Cox, Z Ning, DI Boomsma, E Slagboom

MANUSCRIPT SUBMITTED, 2012

The ubiquitin E3-ligase ARIH2 regulates a muscle-specific expression of PABPN1. H Buijze, SY Anvar, Y Raz, A Venema, SM van der Maarel and V Raz

MANUSCRIPT SUBMITTED, 2012

Extracellular depletion and nuclear entrapment of PCOLCE characterizes muscle pathology in Oculopharyngeal muscular dystrophy. V Raz, S Routledge, E Sterrenburg, SY Anvar, A Venema, BM van der Sluijs, C Trollet, G Dickson, B van Engelen, M Antoniou and SM van der Maarel

MANUSCRIPT SUBMITTED, 2012

Poly(A) binding protein nuclear 1 (PABPN1) levels affect alternative polyadenylation. E de Klerk, A Venema, SY Anvar, JJ Goeman, JT den Dunnen, SM van der Maarel, V Raz, PAC ’t Hoen

MANUSCRIPT SUBMITTED, 2012

Skeletal muscle aging is regulated by PABPN1 expression level. SY Anvar, Y Raz, A Venema, MLR van ’t Hoff, M Gheorghe, JJ Goeman, B van der Sluijs, B van Engelen, M Snoeck, J Vissing, SM van der Maarel, PAC ’t Hoen and V Raz

PLOS COMPUTATIONAL BIOLOGY, 2011 Nov;7(11):e1002258. doi: 10.1371/journal.pcbi.1002258

Interspecies translation of disease networks increases robustness and predictive accuracy. SY Anvar, A Tucker, V Vinciotti, A Venema, GJ van Ommen, SM van der Maarel, V Raz and PAC ’t Hoen

AMERICAN JOURNAL OF PATHOLOGY, 2011 Oct;179(4):1988-2000

Modeling Oculopharyngeal Muscular Dystrophy in Myotube Cultures Reveals Reduced Accumulation of Soluble Mutant PABPN1 Protein. V Raz, S Routledge, A Venema, H Buijze, E van der Wal, SY Anvar, KR Straasheijm, R Klooste, M Antoniou and SM van der Maarel

SKELETAL MUSCLE, 2011 Apr 4;1(1):15. doi:10.1186/2044-5040-1-15
Deregulation of the ubiquitin-proteasome system is the predominant molecular pathology in OPMD animal models and patients. SY Anvar, PAC 't Hoen, A Venema, B van der Sluijs, B van Engelen, M Snoeck, J Vissing, C Trollet, G Dickson, A Chartier, M Simonelig, GJ van Ommen, SM van der Maarel and V Raz

HUMAN MOLECULAR GENETICS, 2010 Jun 1;19(11):2191-207
Molecular and phenotypic characterization of a mouse model of oculopharyngeal muscular dystrophy reveals severe muscular atrophy restricted to fast glycolytic fibres. C Trollet, SY Anvar, A Venema, IP Hargreaves, K Foster, A Vignaud, A Ferry, E Negroni, C Hourde, MA Baraibar, PAC 't Hoen, JE Davies, DC Rubinsztein, SJ Heales, V Mouly, SM van der Maarel, G Butler-Browne, V Raz and G Dickson

BMC BIOINFORMATICS, 2010 Jan 15;11:32; doi:10.1186/1471-2105-11-32
The identification of informative genes from multiple datasets with increasing complexity. SY Anvar, PAC 't Hoen and A Tucker

> “Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.” **Winston Churchill**

acknowledgement

Now, it is so that I conclude my thesis by acknowledging the immense efforts and help from a large group of people that made this journey possible. I do not know, cannot guess and have no way of finding out the significance of these contributions. I do know, however, that without them this work would not have been completed. Hereby, I wish to express my gratitude to those that made this dream a reality.

“The process of scientific discovery is, in fact, a continual flight from wonder.” **Albert Einstein**

First, none of this would have been possible without Silvère, Peter-Bram, Vered, and Allan. I wish to thank them for giving me the opportunity to work on an interdisciplinary and challenging project. Their fascination with the newest developments in the field of systems biology has made this journey an ultimate learning and rewarding experience which I enjoyed greatly.

“Knowledge is knowing a tomato is a fruit; Wisdom is not putting it in a fruit salad.”

I would have been lost in my project without the guidance of Peter-Bram, Vered, and Allan. Their genuine commitment, hard work and broad area of knowledge inspired me to live each day as if a new fascinating project had just begun. Nevertheless, to paraphrase Darwin, a bioinformatician is a blind man in a dark room looking for a black cat which isn’t there. As curious I might be to try new ideas, it has always been their questions and criticism that enlightened me to find my way around. Thank you for your kindness, friendship, and encouragements.

“I would rather walk with a friend in the dark, than alone in the light.” **Helen Keller**

I would like to thank my OPMD family, Andrea, Nisha, Eleonora, Hellen, Erik, Daphne, and Merel, not only for being colleagues but as for being good friends, fun times, and their generous support throughout this period. I would like to thank Andrea for playing a key part in our group and for her important contribution on almost every manuscript we wrote. Special thanks to the Frants group for all your help, fun coffee breaks, and interesting conversations. I would like to thank Antoine, Peter T., Dwi, Antonietta, and Laura for friendship, fun times, your enthusiasm, and helpful discussions. A very especial thank to Anita and Babs for their support throughout all the years. Finally, thanks to all others in the Human Genetics department for your help, contributions, and friendship.

“No! Bioinformaticians are not lazy or boring. They are just protective of their seats.”

From the bioinformatics-core, thanks to Judith, Maarten, Herman, Eleni, Irina, Jelle, Erik, Marco, Kostas, and Harish for helpful discussions, friendship, and despite what is perceived by others, thanks for good times. From the LGTC, thanks to Michel, Jeroen, Martijn, Michael, Henk, Ken, Sophie, Yavuz, Arnoud, Rolf for your support and giving me the chance to work in such a great environment. From the Brunel University, thanks Veronica for your collaborative work and constructive discussions.

“Best friends are those who, when you show up at their door with a dead body, say nothing, grab a shovel, and follow you.”

I’d like to thank my friends for all the memories, good times, laughs, and for your precious friendship. Thanks Malin, Elisenda, Jelrik, and Roel for fun times and unforgettable memories. I’ll cherish them forever. Thanks Shayan, Saba, Kin, James, Ravi, and Ali E. for your friendship and good times we had back in London. Thanks to Caroline, Rob, Liz, and Reuben for your friendship. Thank you for empowering me to make my dreams come true and to stand by me.

“The only true wisdom is in knowing you know nothing.” **Socrates**

First, I would like to thank Ali for our nightlong debates, constructive discussions, and his support. For me, the biggest breakthroughs that helped me decide to go on this path were due to those penetrating conversations. I have to thank Saied, the great, for making this journey a joyful and fun experience. Your sincere friendship helped me through thick and thin.

During all the years of fun, stress, and hard work, I have to thank Fleur for standing by me and making my life truly brilliant. Your patience with me and your unconditional support and compassion made my life a living dream. Thank you to Oma, Jaap Jan, Inge, and Lisa for welcoming me into your lives and making me feel at home.

Finally, thank you to all of my family. None of this would be possible without the help and support of my siblings. Special thanks to Checkad for playing a big part in enabling me to move forward. No word can express my gratitude towards my Mom and Dad. Your passion for knowledge, your encouraging attitude towards one’s betterment, and your ambitious dreams inspired me the most. Your true devotion and lasting belief in me made a dream a reality. For that I love you and I am most grateful. Thank you.

Seyed Yahya Anvar