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

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

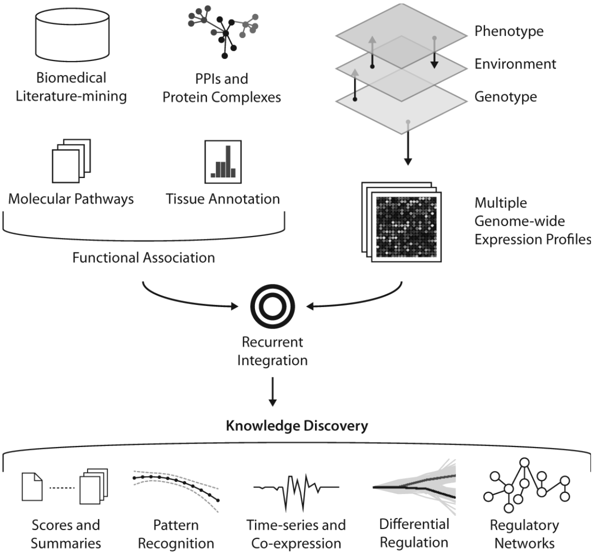


Figure 1 – Schematic illustration of data integration. The recurrent integration of biological data (genome, transcriptome, phenome, and environment) requires special efforts in utilising data sources, protein-protein interaction (PPI) networks and protein complexes, biomedical literature, etc. The proper 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).

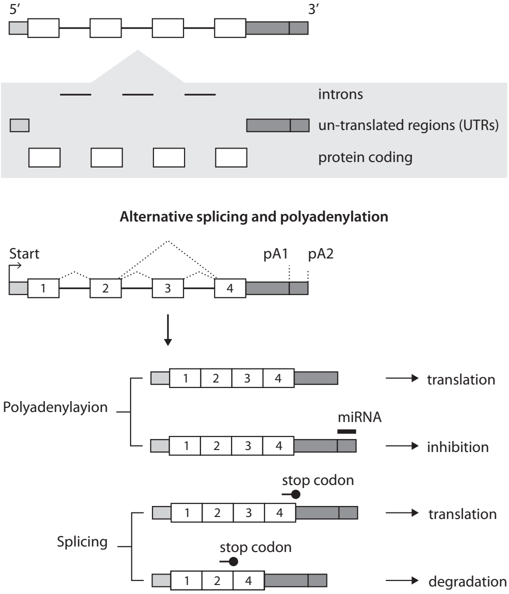


Figure 2 – Schematic overview of RNA processing and its regulation. A single gene can generate pre-mRNAs that are alternatively processed to generate a diverse set of mature mRNAs. These isoforms can differ in inclusion of exons (alternative splicing) and the polyadenylation sites in the 3' UTR (alternative 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).

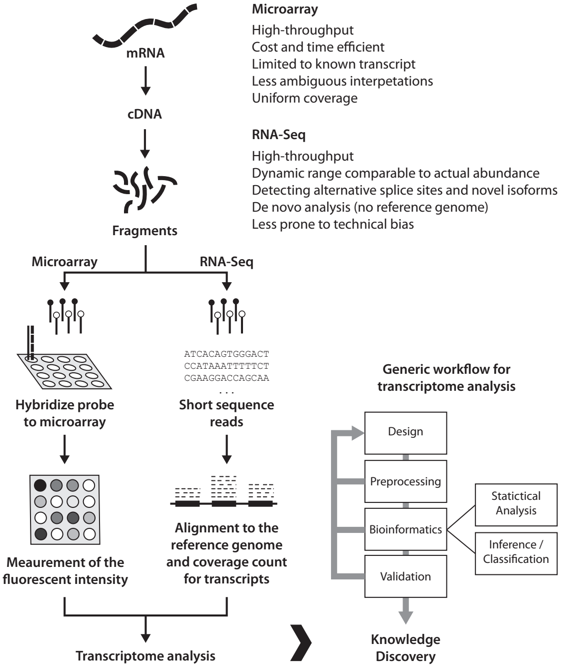

mRNA
cDNA
Fragments
Microarray
RNA-Seq
ATCACAGTGGGACT
CCATAAATTTTTCT
CGAAGGACCAGCAA
...
Hybridize probe to microarray
Short sequence reads
Meaurement of the fluorescent intensity
Alignment to the reference genome and coverage count for transcripts
Transcriptome analysis
Microarray
High-throughput
Cost 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

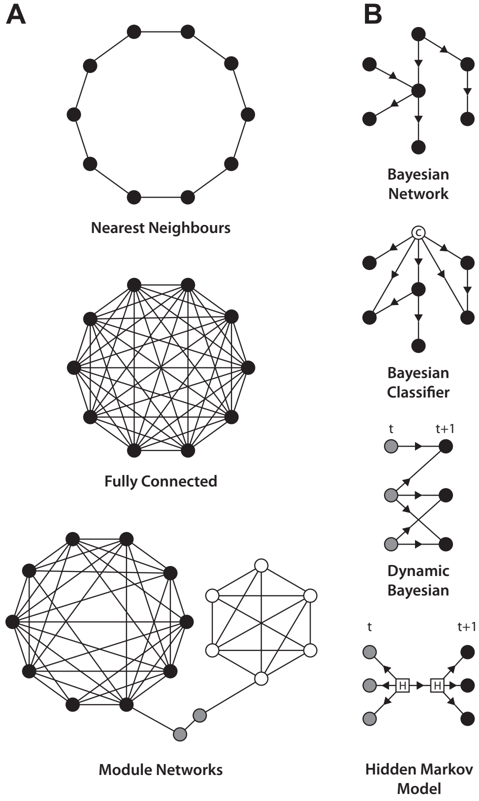


Figure 4 – Schematic illustration of biological networks. A) Co-expression networks can be constructed under various constraints and settings. A cluster of ten nodes can be interconnected on the basis of their nearest neighbours, depicted as a ring. Fully connected networks of ten nodes represent a cluster 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).

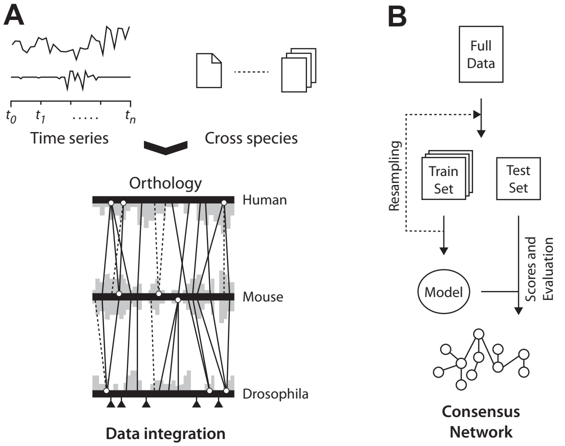
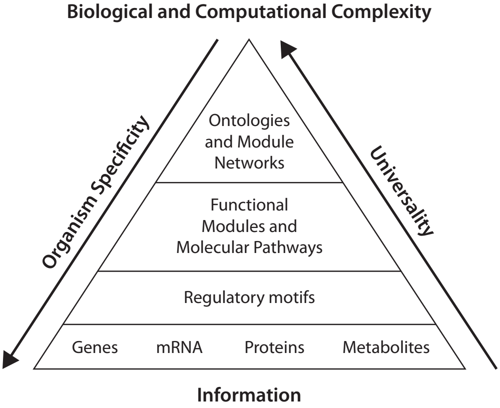


Figure 5 – Bayesian regulatory networks in computational biology. A) Interspecies (or inter-platform) integration can be achieved by taking into account the many-to-many relationships of orthologue genes/transcripts (depicted by circles). Depending on the technology used for generating biological data, 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**).

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.


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

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-

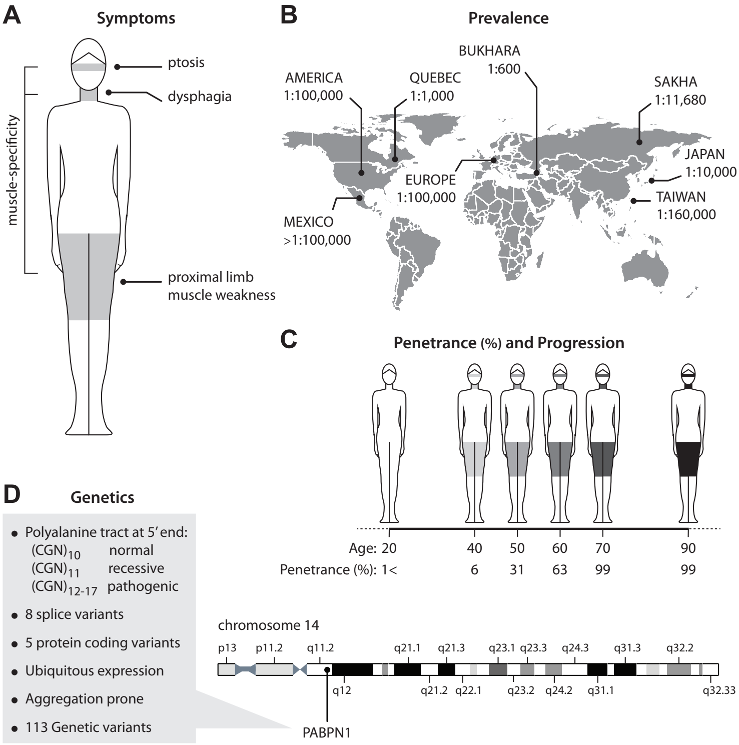


Figure 7 – Schematic characterisation of oculopharyngeal muscular dystrophy. A) OPMD symptoms are mainly restricted to skeletal muscles. **B)** Divers prevalence rates of OPMD estimated in different populations. Worldwide prevalence is estimated to be 1:100,000. **C)** Penetrance (%) and progression 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.