

Cover Page



Universiteit Leiden



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

Interspecies translation of disease networks increases robustness and predictive accuracy

Seyed Yahya Anvar^{1,*}, Allan Tucker², Veronica Vinciotti², Andrea Venema¹, Gert-Jan B. van Ommen¹, Silvère M. van der Maarel¹, Vered Raz¹ and Peter A.C. 't Hoen¹

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.

Species	Tissue	Samples	Age /Time-Point	GEO Accession	Cross-Validation			Number of Networks		
					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.

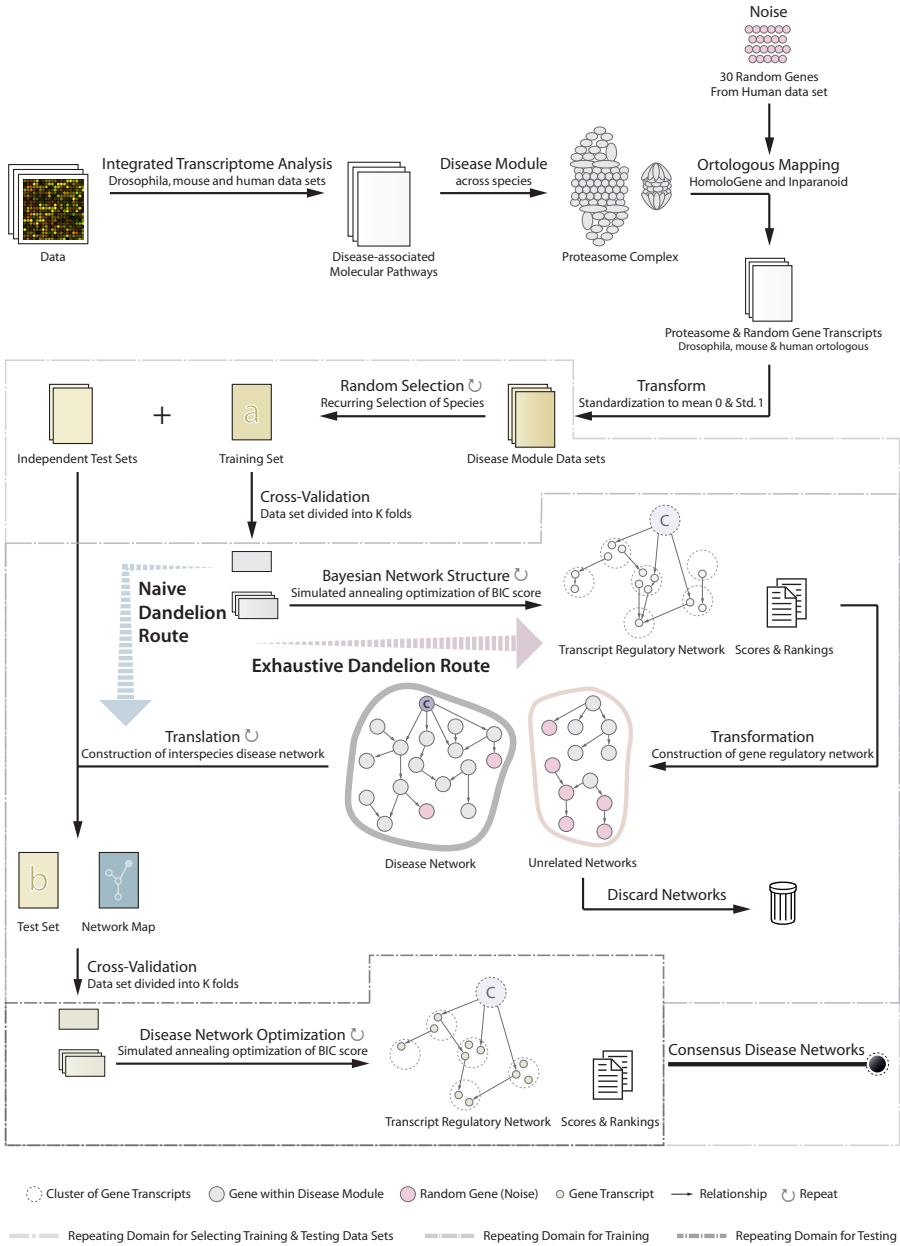


Figure 1 – Schematic overview of the Dandelion algorithm for disease network analysis. The Dandelion algorithm involves three recurring stages of training and independent testing regime with the use of multiple datasets derived from different species. In the first step, disease modules are defined as 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 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 rela-

tionships 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 of the selected organism. Furthermore, the series of $Species_{test_1} \dots Species_{test_M}$ and $test_{folds_1} \dots test_{folds_M}$ 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 μ L 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 $\Delta\Delta$ CT 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.

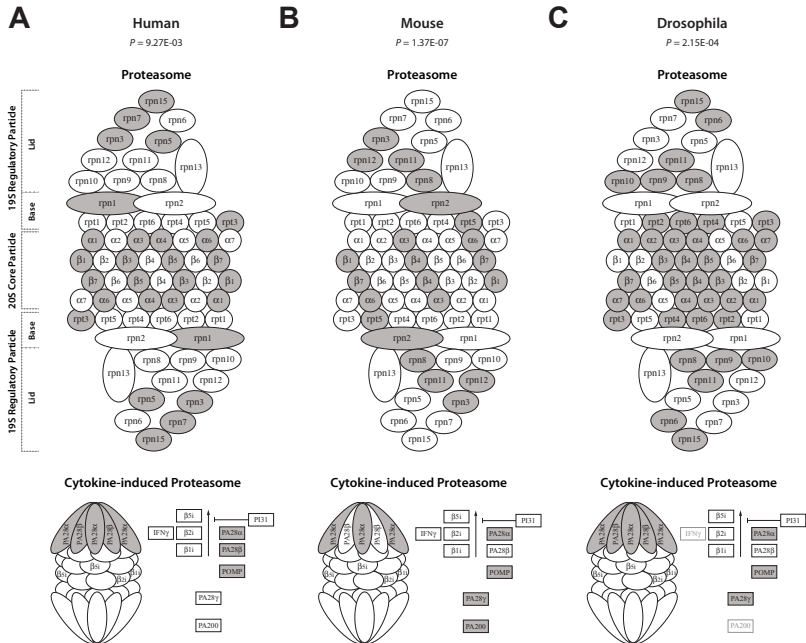


Figure 2 – OPMD-deregulation across different subunits of the proteasome in different species. There are widespread differences in gene expression (depicted in dark colors) between OPMD and control in the different functional subunits of proteasome and immunoproteasome in human (A), mouse (B) and *Drosophila* (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).

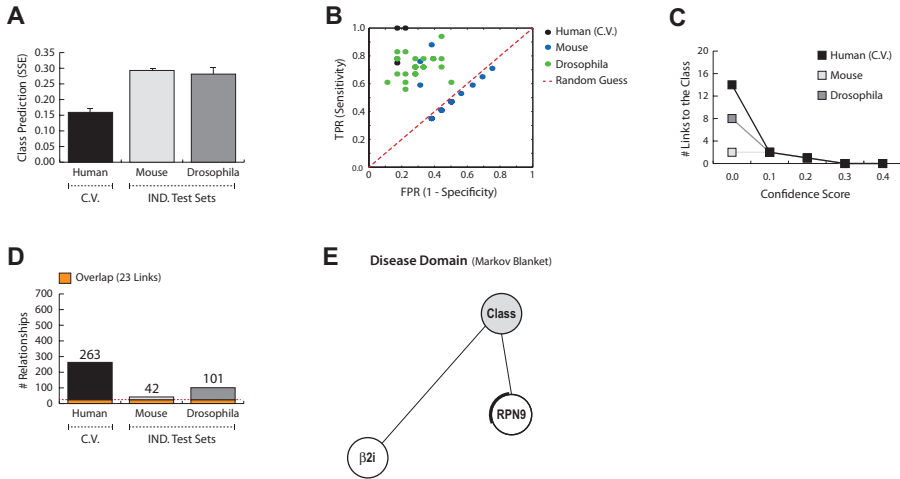


Figure 3 – Performance of the naïve Dandelion algorithm on constructing disease networks that are learnt on human and evaluated on human, mouse and *Drosophila* datasets. A) The average Sum of Squared Error (SSE) for prediction of the disease phenotype (OPMD vs. control) given the gene expression profiles 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

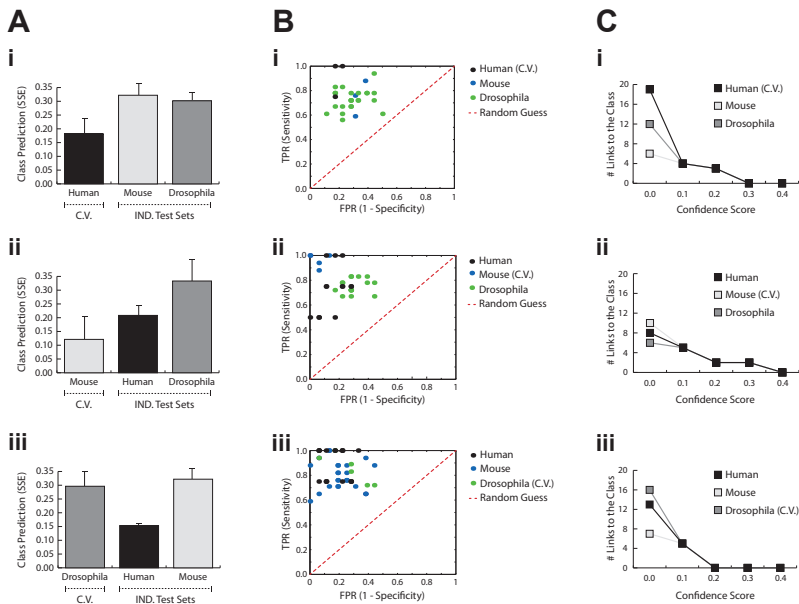


Figure 4 – Performance of the exhaustive Dandelion algorithm. A) The average Sum of Squared Error (SSE) for prediction of the disease phenotype (OPMD vs. control) given the gene expression profiles within the disease networks learnt on human (i), mouse (ii), or *Drosophila* (iii). The cross-validation 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.

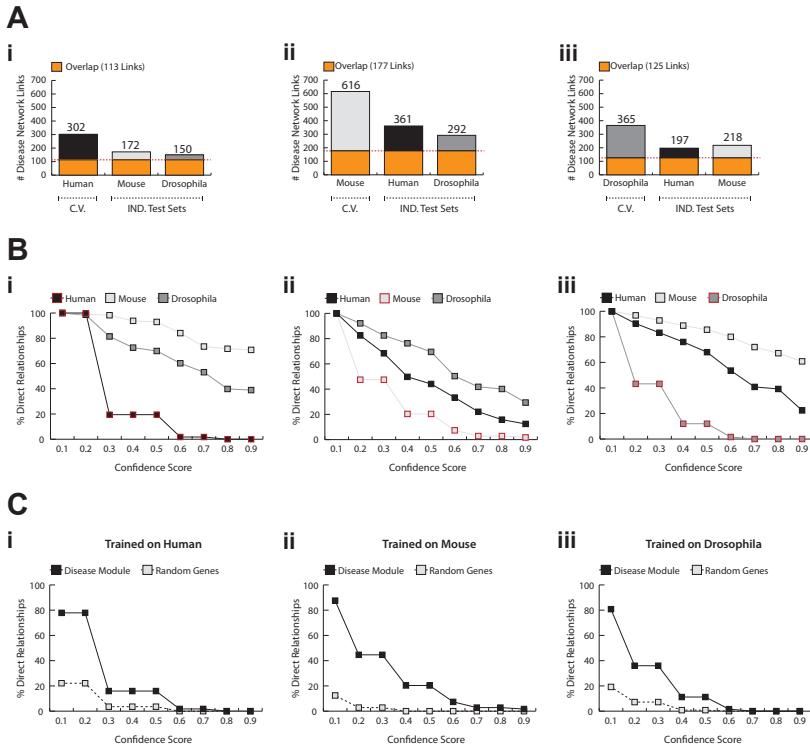


Figure 5 – Translatability and robustness of interspecies disease networks. A) The number of links that were found during interspecies translation and optimization of the disease networks per individual datasets. The red dash-line depicts the number and fraction of links that can be found in all species 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

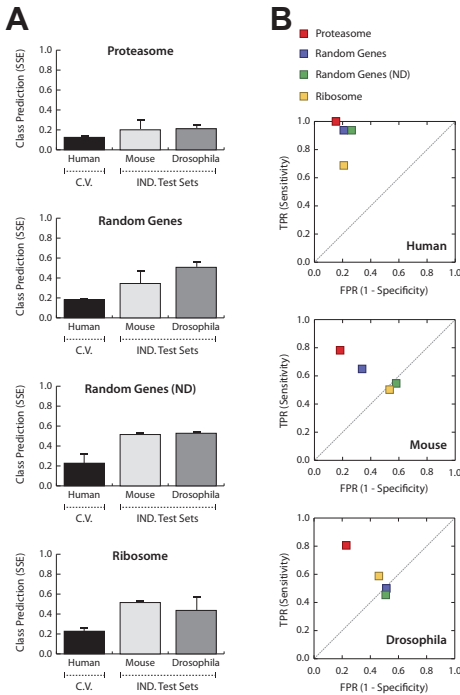


Figure 6 – Specificity of the proteasome towards prediction of disease states. A) The average Sum of Squared Error (SSE) for prediction of the disease phenotype (OPMD vs. control) given the gene expression profiles within the constructed networks learnt on the proteasome, 100 random genes, 70 not-deregulated 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

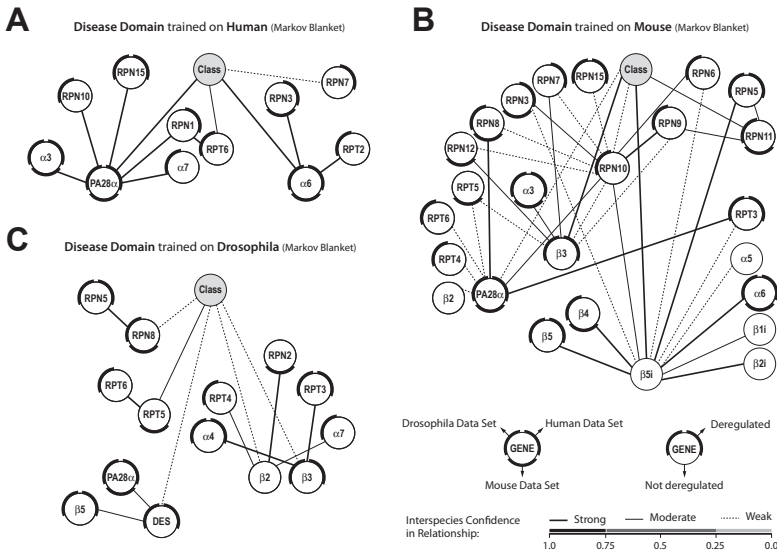


Figure 7 – Interspecies disease domains. These interspecies class network structures are learnt on human (A), mouse (B), or *Drosophila* (C) dataset and optimized across species. Class network structures are presented according to Markov blanket criteria. Nodes represent genes. The outer ring reflects 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

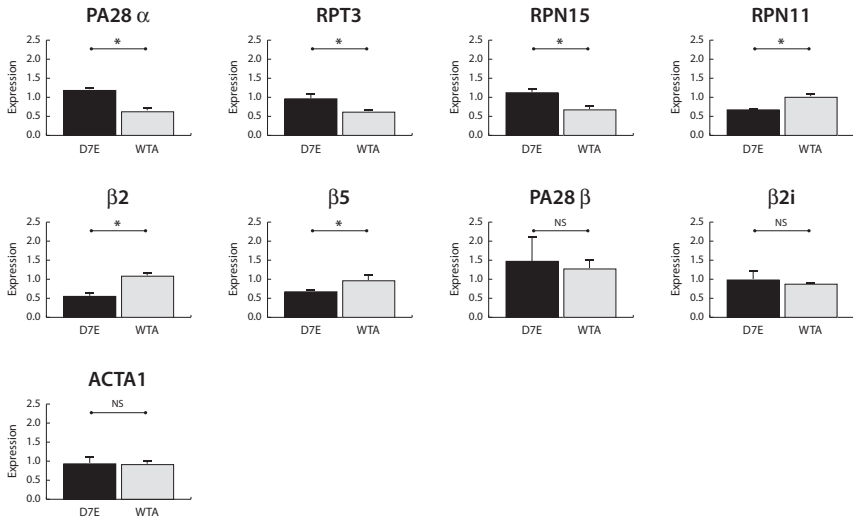


Figure 8 –Validation of differential expression of disease associated genes in an unseen disease model. Results from qPCR experiments measuring differences in gene expression between control cells (WTA, N=3 independent cultures) and cells expressing the OPMD-associated PABPN1 with expanded repeat (D7E, 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 (*PA28 α* , *RPT3*, *RPN15*, *RPN11*, β 2, and β 5) showed significant differential expression in an unseen IM2 cell model (Figure 8). *PA28 α* 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, *PA28 β* , 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 *PA28 β* is not deregulated between the D7E and WTA cells (Figure 8). Next, we assessed the expression of the β 2i, 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; Peèr 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 PA28 α , 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. PA28 α 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 PA28 β 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 β 2 and β 5, 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 β 2 and β 5 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., Verris,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., Linal,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-validated 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	<p>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.</p> $SSE_g = \sum_{i=1}^n (\text{measured value}_{g,i} - \text{predicted value}_{g,i})^2$
Sensitivity	<p>The probability of accurate prediction of cases with the disease-associated phenotype.</p> $\text{Sensitivity} = \frac{\text{number of True cases}^{\text{disease}}}{\text{total number of True cases}^{\text{disease}}}$
Specificity	<p>The probability of accurate prediction of control cases without the disease-associated phenotype.</p> $\text{Specificity} = \frac{\text{number of True cases}^{\text{control}}}{\text{total number of True cases}^{\text{control}}}$
Confidence Score	<p>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.</p> <p>For the training set (species A):</p> $\text{Confidence Score} = \frac{\text{number of times a link is found} (n_{\text{species A}})}{\text{total number of constructed networks on A}}$ <p>For the independent test set (species B):</p> $\text{Confidence Score} = \frac{\text{number of times a link is found} (n_{\text{species B}})}{n_{\text{species A}} \times \text{total number of constructed networks on B}}$
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.
<hr/>	
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.
<hr/>	
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.
<hr/>	
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

$t = t \times c$

End while

Output: $result$

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

Input: $Species_{train}$ { $Species_{test\ 1}$, ..., $Species_{test\ M}$ }, $train_{folds}$ { $test_{folds\ 1}$, ..., $test_{folds\ M}$ }, $exhaustive_{T,F}$

For $k = 1$ **to** $train_{folds}$

Learn $intraspeciesTranscript_{bn}$ using **Algorithm 1** on training folds of $Species_{train}$

Score $Species_{train}$ { $Nodes_{SSE}$, $Nodes_{STD}$, $Links_{Confidence}$ }

If $exhaustive = true$

Transform $intraspeciesTranscript_{bn}$ **to** $intraspeciesGene_{bn}$

End if

Assess Disease Connection

If $intraspeciesGene_{bn}$ is not connected to *disease node* **then**

Drop $intraspeciesGene_{bn}$

Else

Translate $intraspeciesGene_{bn}$ **to** $networkMap$

For $i = 1$ **to** M

Optimize and **Test** $networkMap$ in $Species_{test\ i}$ using **Algorithm 1**

Score $Species_{test\ i}$ { $Nodes_{SSE}$, $Nodes_{STD}$, $Links_{Confidence}$ }

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	CACNA11	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

CHAPTER FIVE

PSMC5	LOC400831	NOC4L	LOC652610	OGG1	HDX	RPL15	RPS21
PSMC6	HMGN4	PLD3	LOC646699	TAF9B	APOL3	RPL17	RPS23
PSMC3	TSSK4	LOC648974	KNDC1	LOC653421	CNOT4	RPL18	RPS24
PSMC4	RTKN2	GTPBP8	DACT3	LOC441347	PFAS	RPL18A	RPS25
PSMA6	RXRA	LIF	FLJ16369	FRMPD2	MAP3K14	RPL19	RPS26
PSMA2	MYL5	LOC440104	VIPR1	HSCB		RPL21	RPS27
PSMA4	UBTD1	WAC	COPS8	CHD1		RPL22	RPS27A
PSMA8	OR1J4	KALRN	NIF3L1	LOC645781		RPL23A	RPS28
PSMA7	TRAPPC5	UNC93A	PPAP2C	LOC729446		RPL24	RPS29
PSMA5	ADAM20	IFNAR1	LOC644431	FAM129C		RPL26	UBA52
PSMA1		NMT1	TCTE3	FAM90A15		RPL27	RPL14
PSMA3		LOC652750	TTF2	C1orf187		RPL30	RPL23
PSMB6		LOC653707	RPS7	HIPK2		RPL27A	RPL35
PSMB7		SLC26A9	ITGA8	XKR3		RPL28	RPL13A
PSMB3		ETFDH	CCAR1	RAB2A		RPL29	RPL36
PSMB2		ADAM23	PDCD10	FOXR1		RPL31	MRPL13
PSMB5		FBXO9	LOC651400	CD72		RPL32	RPS27L
PSMB1		LOC643089	CDC42BPG	TRAF4		RPL34	RPL26L1
PSMB4		ATP5D	SP2	NCAN		RPL35A	C15orf15
PSME1		CST6	LOC649432	HRC		RPL36AL	RPL10L
PSME2		RPL11	LOC732093	LOC643577		RPL37	RPL22L1
PSME3		FAM47B	TMEM165	AKR7A2P1		RPL37A	RSL24D1P11
PSME4		LHFPL4	LHCGR	PLK2		RPL38	
POMP		MGC42105	SPAG7	RABL2B		RPL39	
PSMF1		STOX2	INOC1	CLGN		RPL41	
IFNG		FRMD5	OR2T10	LRRC49		RPL36A	
PSMB9		CHL1	DEPDC5	CHORDC1		RPLP0	
PSMB10		UNQ830	ADAD1	KRT18P51		RPLP1	
PSMB8		STCH	LOC339529	OR13G1		RPLP2	
PSMB11		B4GALNT3	FZD9	CCL21		RPS2	
AKR1CL1		SUMO2	CD46	LRFN2		RPS3	
CHRNA5		C20orf30	JARID1B	SLC35A5		RPS3A	
UNC13B		CNIH3	DUX4	RDH12		RPS4X	
DES		DBX2	DPPA4	FAM154B		RPS4Y1	
STT3A		GSTM5P1	YSK4	LOC388948		RPS5	

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 (Psmc14)	CACCTGAACAGCTGGCAATA	GAGCATTGGGAACGAAGAAG
RPN15 (Shfm1)	AGCACGGCTACAAGATGGAG	TGAACCAAAAAGATTAATCAAAACA
RPT3 (Psmc4)	ACCTCAGACCAGAAGCCAGA	CACCACACGGATAAATGCAG
β 2 (Psmb7)	GCACTACCGCTGCCTCACCG	AGGGGTGGTATGCACCCCGAG
β 5 (Psmb5)	CGGTCGCAGCAGCCTCCAAA	GCATACACGGAGCCAGAGCCC
PA28 α (Psmc1)	AAGCCAAGGTGGATGTGTTT	GGGTACTGGGATGTCCAATG
PA28 β (Psmc2)	CCTGGAGAGTGAAAGCGAAA	GTCATCAGCCTCCTGGAAAA
β 2i (Psmb10)	ATTTGCTCCTGGAACACAC	CCACTTCATTCCACCTCCAT
ACTA1 (Acta1)	CGAGGTATCCTGACCCTGAA	AGGTGTGGTGCCAGATCTTC
mHPRT	CGTCGTGATTAGCGATGATG	TTTTCCAAATCCTCGGCATA

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

Gene A	Gene B	Train Set	Interspecies Confidence		Human		Mouse		Drosophila	
			Correlation Score	P-value	Correlation Score	P-value	Correlation Score	P-value	Correlation Score	P-value
PA28 α	RPN1	Human	Strong	2.80E-03	0.6059	0.8650	8.50E-11	-0.3762	2.37E-02	
PA28 α	RPN15	Human	Strong	8.37E-01	-0.0466	0.7521	4.50E-07	0.4801	3.03E-03	
PA28 α	RPN8	Mouse	Strong	7.57E-01	-0.0670	0.9245	1.58E-14	0.4658	4.20E-03	
PA28 α	RPT3	Mouse	Strong	3.26E-03	0.5984	0.7944	3.42E-08	0.4491	6.00E-03	
PA28 α	α 3	Human	Strong	1.24E-02	-0.5234	0.8988	1.24E-12	0.5498	5.14E-04	
RPN10	RPN3	Mouse	Moderate	8.72E-03	0.5450	0.6400	6.07E-05	0.5503	5.06E-04	
RPN10	RPN6	Mouse	Moderate	1.68E-03	0.6300	0.8634	1.00E-10	0.8180	1.12E-09	
RPN10	RPN9	Mouse	Strong	1.29E-01	0.3338	0.8760	2.46E-11	0.8545	3.36E-11	
β 3	RPN12	Mouse	Moderate	2.98E-02	0.4635	0.8653	8.15E-11	0.5090	1.52E-03	
β 3	RPN7	Mouse	Moderate	2.73E-01	-0.2445	0.9386	7.17E-16	0.6361	3.05E-05	
β 3	RPT3	Drosophila	Strong	4.72E-02	0.4274	0.8652	8.32E-11	0.5855	1.76E-04	
β 3	α 3	Mouse	Moderate	6.87E-01	0.0910	0.9456	1.15E-16	0.8367	2.07E-10	
β 3	α 4	Drosophila	Strong	2.91E-01	0.2359	0.9378	8.65E-16	0.7959	6.56E-09	
β 5i	α 6	Mouse	Strong	3.28E-03	-0.5980	0.6593	3.01E-05	0.6393	2.70E-05	
β 5i	β 1i	Mouse	Moderate	6.00E-04	0.6729	0.9053	4.63E-13	-0.1952	2.54E-01	
β 5i	β 4	Mouse	Strong	3.76E-01	0.1983	0.7679	1.83E-07	0.6416	2.48E-05	

