Article:

Davidsen, PK, Turan, N, Egginton, S et al. (1 more author) (2016) Multi-level functional genomics data integration as a tool for understanding physiology: A network biology perspective. Journal of Applied Physiology, 120 (3). pp. 297-309. ISSN 8750-7587

https://doi.org/10.1152/japplphysiol.01110.2014

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

Peter K. Davidsen[1], Nil Turan[2], Stuart Egginton[3] and Francesco Falciani[1*]

[1]Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool, UK.

[2]School of Biosciences, University of Birmingham, Birmingham, UK.

[3]School of Biomedical Sciences, Faculty of Biological Sciences, University of Leeds, U.K.

[*]Correspondent author

Contact Information:

Prof. Francesco Falciani
Centre for Computational Biology and Modelling, Institute for Integrative Biology
University of Liverpool
Crown Street
L69 7ZB
Liverpool, UK
Tel: +44-151-795-4558
Email: f.falciani@liverpool.ac.uk

Running Head: Multi-level network integration to understand physiology.

ABSTRACT

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

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

INTRODUCTION

Modelling in physiological sciences

Physiology has evolved as a series of sub-disciplines attempting to understand organismal function as a combination of interacting components and systems. The last decade or so has witnessed the development of Systems Biology as an investigative approach, and its application in different areas of biology, ranging from engineering/synthetic biology (e.g. design of bacterial strains with improved properties) to health sciences (e.g. disease biomarker identification). Despite the lack of a concise definition acceptable to the majority of the community (30, 32), Systems Biology is frequently understood to be the study of complex regulatory interactions in biological systems using a holistic approach. This is often achieved by integrating different experimental approaches within the conceptual framework of a computational model (i.e. a mathematical representation of a system that allows simulation of its behaviour). Physiology is probably one of the few research areas in biological sciences that have traditionally adopted such an approach. It has long sought to understand the behaviour of complex biological processes and cellular systems using an integrative approach, and has extensively adopted mathematical modelling in its tool set. Classical examples include August Krogh's tissue cylinder model of oxygen transport to skeletal muscle (33), and Huxley's two-state cross-bridge model of muscle contraction (26), which are still used by investigators today. Indeed, this shows that using modelling to study a system as a whole has been a key component of physiology from its early days.

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

The advent of Functional Genomics: a challenge for physiological modelling

It is now clear that much of the complex mammalian physiology or pathophysiology cannot be understood in sufficient detail through a reductionist approach alone. Although this approach has proved valuable in explaining broad phenomena and individual mechanisms, linking multiple mechanisms and effects has proved challenging. For example, a disease phenotype is rarely caused by a single dysfunctional gene or protein. Instead, genetic variability, epigenetic modifications, post-transcriptional regulation mechanisms etc. all act in concert to determine a specific high-level phenotypic response (43). The potential for such complex interaction makes data interpretation much more complicated than originally envisioned, highlighting the need to move away from the widespread 'candidate gene' approach (39).

Triggered by the advent of genome sequencing, inspired by the Human Genome Project, dramatic technological advances within the last decade or so have led to increased throughput in genome-wide molecular analyses (i.e. genomics, epigenomics, transcriptomics, proteomics, metabolomics). The comprehensive data acquisition tools developed to cope with large datasets have allowed investigators to determine the molecular state of cells, tissues or even entire organs in a single experiment. Such cost-effective omics approaches are now becoming prevalent in biological and medical research, and consequently have been responsible for the generation of an incredibly large amount of multivariate molecular data. A large proportion of this data is available in the public domain via different online databases (e.g. NCBI Gene Expression Omnibus (5), EBI ArrayExpress (7), and PRIDE (29)).

For example, mRNA microarray technology and more recently mRNA sequencing, has provided insight into the transcriptional response of skeletal muscle to prolonged endurance exercise training, highlighting a pronounced inter-individual variation at the molecular level that is consistent with the heterogeneous response observed in a population of individuals at the physiology level (31, 59). Statistical models built to explain such variation as a function of gene expression data can be exploited to identify underlying mechanisms controlling tissue homeostasis. The transcriptional signatures identified in such studies likely explain, at least in part, why some people show great improvements in aerobic capacity (VO_{2max}) whereas others only experience smaller benefits, despite completing the same supervised exercise training program. Another example of applying omics technology to better understand human physiology concerns the quantification of individual levels of different proteins in health and disease; by use of proteomics methodology, Holloway et al. (24) were the first to investigate adaptations in human muscle protein content to long-term exercise training on a large scale.

While such omics-based studies hint at the potential of a data-driven approach, they also illustrate the difficulty in deriving conceptual models underlying the essential mechanisms regulating physiology, as most are restricted to only one aspect of regulation. Perhaps surprisingly, the exponential growth in publicly available omics data (34, 37) has not resulted in a paradigm shift in our understanding of biology. The main reason is the continuing challenge of integrating multivariate datasets spanning multiple organization levels in a way that allows the identification of discrete, small biomolecular networks that are truly important in the context of a specific biological response (47). Such a task cannot be achieved simply using unaided human interpretation. Rather, complex computational techniques are needed that are able to integrate and automatically 'learn' the structure of a biological system. Such a modelling framework is very different from what physiological sciences have traditionally employed.

Towards data-driven predictive biology

Although the modelling approach traditionally used by physiologists has been extremely successful, it suffers from severe limitations when challenged with extensive omics data. For example, physiological modelling relies to various degrees on a mechanistic understanding of the biological system of interest (16), which automatically limits the number of components that can be included due to gaps in our current knowledge (19, 47). Moreover, estimation of model parameters, which is usually a challenging task because of experimental limitations (e.g. due to limited amount and quality of data), makes the approach difficult to scale up to a larger number of components and their interactions. Perhaps the most comprehensive example to date is modelling the cardiac cycle based on ion channel kinetics (44).

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

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

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

COMPUATIONAL APPROACHES FOR THE ANALYSIS OF COMPLEX DATASETS

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

[Figure 1 to be inserted here]

The first approach is biomarker discovery (**Figure 1A**), which perhaps is most widely used in the analysis of functional genomics datasets. Here the objective is to identify measurable variables that are predictive of a given outcome (e.g. the response to physical training in a population of individuals). Such measurements can be molecular (e.g. gene expression, protein levels, metabolite concentrations, genetic mutations) and/or more traditional physiological endpoints (e.g. endurance, VO_{2max}). The identification of predictive biomarkers can be achieved by use of univariate and multivariate variable selection strategies that aim to identify the most relevant explanatory measurement(s), while developing a computational model that can accurately predict an outcome (60). Univariate methods will test every variable (e.g. expression of a given gene) on its own, whereas multivariate methods test combinations of variables for their ability to explain a given outcome. Clearly, multivariate approaches better resemble the complex nature of biological networks, and therefore are more likely to provide insights into the mechanisms underlying a complex phenotypic trait. Consistent with this notion, multi-gene biomarkers are often required for robust predictions in independent datasets.

The second approach (**Figure 1B**) consists of ‘reverse engineering’ biomolecular networks from observational data (i.e. infer regulatory interactions between quantified biomolecules based on mathematical principles). Here the overall aim is to reconstruct the underlying structure of interactions between biological molecules profiled using omics tools (ideally from multiple data sources) and rigorous statistics. Such a network inference framework can be achieved by applying a multitude of approaches with varying underlying data assumptions and modelling principles, including ordinary differential-equation (ODE)-based methods (3), probabilistic modelling techniques (e.g. Bayesian theory models) (42, 64), state-space representation models (23), and correlation-based methods. Note, while the first three approaches are able to infer directed networks, their capability is currently limited to inferring smaller networks with few variables due to increased computational complexity than possible with correlation approaches.

Importantly, this network inference part may potentially benefit from a biomarker discovery phase, since it has been shown that identified predictive variables are more likely to be directly controlling important physiological processes, and therefore are good candidates to include in a network (47). Similarly, whole networks can be used as an input for biomarker discovery procedures. It has been shown that often the overall 'activity' of a biological network (e.g. a specific signalling pathway) is a better predictor than a few key individual genes, proteins and/or metabolites. This implies that in the coming years predictive biomarkers are more likely to consist of a relatively large panel of measurements, possibly spanning multiple levels of complexity within a pathway. Current omics platforms are experiencing a rapid development as well as drop in costs, making routine collection of large datasets a feasible option. Once a robust biological network has been inferred this may serve as a good basis for developing a more conventional modelling approach to provide explanations for observed phenomena that requires a mechanistic understanding of the system (**Figure 1C**).

Finally, multiple computational models that initially were developed independently can be integrated into a larger and more complex models, which allow responses to physiological/pathological challenges to be simulated, thus integrating effects across multiple organs and/or pathways. These complex models are often referred to as decision support systems because of their potential to provide information about the expected outcome of a therapeutic intervention (**Figure 1D**).

Several large international projects aiming at the development of such technology into Systems Medicine integrated frameworks have been established so far, e.g. the Virtual Physiological Human (VPH) project funded by the European Commission 7th Framework Programme, which aims to aid clinically relevant research by establishing a framework for handling and integrating various mechanistic models spanning different levels of organizational complexity (ranging from molecular components to organ function). By unifying the modelling languages employed across the different mathematical models included, parameters of a particular model in the hierarchy can be processed by other appropriate models at a lower hierarchical level. These global initiatives should be considered long-term goals, aiming at understanding human physiology quantitatively as a dynamic system. Developing a comprehensive model of a biological system requires integrating mechanistic and probabilistic inferences. The mathematics for performing such a task is in its infancy, and more development is needed. However, a successful example is illustrated by the anatomically based model of human heart ventricles (44). In the following sections we aim to provide an overview of some of the methodologies that can be used to infer biomolecular networks, as well as introduce one particular approach we have found useful in our research.

Inference of biological networks from observational data

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

Static vs. dynamic networks

Biological networks can be reconstructed from two different types of experimental studies: either cross-sectional, e.g. representing a population of individuals at a given time (i.e. steady-state measurements following an experimental perturbation), or prospective, where the experimental data is available across a defined time-course. In reverse engineering, statistical inference of biological causality is an important goal (56). A simple example of causality could, for example, be a transcription factor regulating the expression of several target genes. Since determining cause and effects implies a direction (i.e. the cause precedes the effect), inference of causality from cross-sectional studies presents a challenge due to their static nature, one that is less difficult when a time-course is available. However, it must be stressed that both approaches are often used in combination to, for example, integrate clinical cross-sectional studies (thereby providing the researcher with a static network representation) and experimental intervention studies that can provide dynamic (prospective) models of the process being studied. At present, most of the developed techniques infer regulatory networks without any causality information (likely due to the scarcity of time-course datasets due to their higher costs). However, a small number of causality detection techniques have been proposed in the literature such as dynamic Bayesian networks (48) and Granger causality (46). It is also important to point out that true time-course datasets can only be developed when the sequence of events is measured within the same cells/tissues. This is for example achieved with imaging techniques that require complex molecular probes, and can typically be only applied to measure a relatively small number of system components (14). Omics technologies unfortunately are disruptive, so time course data derived using these approaches are in fact a sequence of independent snapshots, which clearly limits the potential use of dynamical modelling tools.

A primer for network inference methods

The simplest method for inferring statistical relationships between experimental variables is computing the pairwise correlation coefficient across a large collection of heterogeneous samples (8). Usually such an approach is not able to identify complex non-linear dependencies, and does not discriminate between direct and indirect connections. More complex methods, such as the mutual information (MI) based Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (38), also aim at establishing a statistical relationship between pairs of variables but have a stronger theoretical foundation. Because of the added mathematical complexity they can capture a broader range of biologically relevant dependencies between variables including non-linear, non-monotonic relationships; importantly, they can distinguish between direct and indirect relationships. ARACNE is a free tool for which a Java-based graphical user interface exists; hence investigators do not need any programming skills in order to use the software.

ARACNE relies on estimating the probability that a variable (e.g. the expression of a gene or a protein) assumes a certain ‘state’ (i.e. abundance) given the state of another biomolecule (conditional probability). A number of alternative MI-based implementations have been proposed during the last decade (e.g. Context Likelihood Relatedness (CLR) (13), Minimum Redundancy/Maximum Relevance Networks (MRNET) (41)), which mainly differ by the way inferred indirect relationships (so-called ‘edges’) are removed once the dependencies between all pairs of variables have been mathematically formulated. In such analyses, unwanted indirect interactions occur by default if there is strong correlation between biomolecule 1 and biomolecule 2, and between biomolecule 1 and biomolecule 3 in a three-node clique (i.e. a triplet of connected variables).

An MI value of zero means that there is no dependency (i.e. no information flow) between two variables, whereas an MI value of 1 indicates a perfect association between them, and therefore, a likely strong regulatory interaction between them. For each inferred dependency, a P-value is calculated based on the distribution of MI values between random permutations of the original dataset, thereby allowing the elimination of all non-statistically relevant dependencies by thresholding using an appropriate (user-defined) cut-off level. Importantly, the quality of the inferred interaction network depends on the arbitrarily selected probability cut-off. A small threshold (e.g. P=0.05) gives a high recall (i.e. fraction of true dependencies that could be inferred) but low precision, whereas a higher threshold (e.g. $P=10^{-6}$) yields better precision (i.e. fraction of inferred dependencies that really are in the network) while suffering from a low recall. A further advantage of MI as an information-theoretical measure of dependency between variables concerns its relatively low computational requirements for building an interaction network. Hence, MI is able to handle very large data matrices with thousands of experimental variables, whereas most of the other more advanced techniques mentioned (e.g. Bayesian methods) can only deal with much smaller numbers of variables (<100) because of the high computational complexity. However, in order to infer robust statistical associations based on MI a fairly large sample size is required (> 50-100 biological replicates), due to the required estimation of the (joint) frequency distribution of the connectivity. Interaction networks derived from such reverse engineering methodologies can be visualized and further analysed using various freeware software tools such as Cytoscape (55), Pajak (6), and BioLayout (18). A comprehensive list of visualization tools focused on interaction networks and their web-links has recently been reviewed (17).

Up to now, these information-theoretic approaches have usually been employed on gene expression data only, due to the wealth of such data available. However, as physiologists have known for many decades, biological systems are usually more complex and multi-layered. Indeed, despite some popularist science writing to the contrary, genes on their own are merely permissive elements within biological systems (43). Further, it has been shown that when multiple types of data (e.g. copy number variants, protein or microRNA expression levels) are incorporated in the network inference pipeline, the accuracy of the learned network topology increases (49). Hence, at present there is a call for methodologies that can embed multiple data sources in a single computational framework. Our recent work has focused on methods that are able to handle large-scale, multi-dimensional genomic datasets (9, 21).

Topological analysis of inferred biological networks provides useful biological insight

Up to now we have described some of the most widely used methodologies for inferring regulatory networks. However, an immediate challenge arises in interpreting these often large, complex networks that visually present as a ‘hairball’ (i.e. too dense a collection of connections to comprehend as a whole) (40). A simple solution, although not very objective, is to focus the analysis around a favourite gene(s). In this scenario, the investigator typically examines the manually selected sub-network in order to identify unknown or unexpected biological relationships, which in turn may be used to formulate new hypotheses. Such ‘discovery-led’ science may be useful when there is insufficient information to generate hypothesis.

Alternatively, the topological properties of the network can be used to identify interesting genes and sub-networks that can be interpreted. We and others have demonstrated the existence of a higher-level, modular organization in biological networks (47, 52, 54), i.e. components of biological systems that act in collaboration to carry out specific biological processes. Consequently, several modularization approaches have now been developed to help group subsets of cellular components based on a given property, such as topological structure or functional role. Such decomposition of a large complex network into relatively independent sub-networks (or ‘modules’) has been shown to be an effective way to deduce the underlying structure of the fully connected network containing many hundred variables (so-called ‘nodes’), as each module can then be analysed independently. In addition, studies have demonstrated that such identified network modules can serve as better predictors of a physiological response than the classic biomarker discovery approach (see Figure 1).

In biomolecular interaction networks, as well as sub-networks, nodes have different levels of connectivity (i.e. number of interactions with other nodes). It has been shown that such interaction networks have so-called ‘scale-free’ structure properties, as their node connectivity distribution fits a power law (4). Such a power law degree distribution implies that most of the connections between biomolecules is linked to a small number of highly connected nodes, such that a large proportion of the molecular state of a cell can be explained by a small subset of biomolecules (so-called ‘hub’ nodes; e.g. a transcription factor that regulates many more genes than average). Hence, in biological networks a hub is often assumed to be a key component of a regulatory networks, hence important for the function of a cell/tissue under investigation. This assumption is supported by the fact that random node disruption does not significantly affect the network architecture, whereas deletion of hub nodes leads to a complete breakdown of the network structure (1). Hence, adjusting the spatial position of each node according to its interconnectivity has been shown to be a simple, yet effective way of visualizing large complex interaction networks (57).

More advanced methods to extract information from complex networks exist that aim to identify functional modules (i.e. sub-networks of biomolecules that are linked to the same biological function), e.g. by integrating both physical interactions (i.e. experimentally validated protein-protein interactions) and mRNA expression data (27). In this context, an identified functional module represents a putative multi-protein complex that is transcriptionally regulated in a specific experimental condition (e.g. treatment vs. control). Hence, by considering additional data on a different level of organization, one can potentially infer a clearer composite picture of the underlying biological function.

Finally, in order to generate objective hypotheses about biological processes controlled by a specific hub node or sub-network, functional enrichment analysis can be performed on all its direct neighbours (i.e. all the adjacent nodes that are directly connected to the hub) (25). Such enrichment analysis aims at reducing complexity by defining groups of molecules (represented by gene sets) that share similar biological functions (e.g. a class of adhesion molecules). To accommodate latest advances in knowledge, the different annotation databases used for this purpose (e.g. Gene Ontology (2) and KEGG (45)) are frequently updated by curators. Using software tools like the web-based application DAVID (11) or applications such as BiNGO (36) developed specifically for use with software visualization tools like Cytoscape, one can quickly determine whether any gene sets are statistically over-represented, thus generating hypotheses on the biological processes controlled by those factors outlined above.

CASE STUDY: INFERENCE OF OXYGEN-DEPENDENT PATHWAYS IN SKELETAL MUSCLE

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

This analysis uses two different datasets. The first is a publicly available dataset (GSE27536) representing a cohort of COPD patients and healthy controls matched for age and smoking history (10) (see **Table S1** for subject characteristics), which includes gene expression profiling in vastus lateralis muscle and whole-body physiological variables (e.g. VO_{2max}, minute ventilation, PaO_2) (50)(61). The second dataset represents an unpublished, genome-wide transcriptional response of mouse soleus muscle to a gradual decline in atmospheric oxygen concentration (GSE64076). Using the first dataset, representing the transcriptional state of skeletal muscles in a COPD cohort (**Figure 2A**), we first show how to infer connections between oxygen availability (e.g. VO_2max), oxidative stress (protein carbonylation) and gene expression signatures (**Figure 2A-C**).

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

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

In order to reconstruct an interaction network spanning multiple levels of organization, we have utilised the following strategy that was developed earlier (61).

1. <u>Combining measurements from different data sources</u>

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

Definition of a biological framework for data-driven network inference

The outcome of data-driven reverse engineering of biological networks, in the absence of any biological assumption(s), often provides results that are difficult to interpret due to the large number of inferred significant interactions. Thus, to reduce complexity of the problem, we decided to focus the analysis on the set of physiological parameters and genes encoding for enzymes in the central bioenergetic pathways (i.e. TCA cycle, oxidative phosphorylation, glycolysis) (see **Table S2** for the full list of variables). The latter choice is reasonable considering the paramount importance of these molecular pathways in skeletal muscle adaptation. The overall strategy is therefore to identify biomolecules that are highly correlated (based on MI) with biologically important experimental variables. Such a focused analysis will generate multiple network modules of interacting biomolecules, each with a bioenergetic hub gene or physiological measurement at its centre. Two modules will be linked together if a specific gene is statistically linked to both hubs.

2. Reverse engineering.

In order to infer robust regulatory relationships between variables in the integrated multi-level dataset, we used the ARACNE algorithm. This choice was based on the large number of measured variables to be considered by the mathematical framework.

By combining all genes expressed in human skeletal muscle (>10,000 mRNAs) with the list of physiological variables we far exceed the number of variables that can be handled by more advanced network inference methods (e.g. Bayesian methods). Hence, we infer a static network without any obvious hierarchical organization. The result of an ARACNE run is an 'adjacency matrix' containing MI values for all pairwise interactions above the specified MI threshold, which can be visualized automatically in Cytoscape.

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

3. Network visualization

Data visualized as a network are often easier to interpret than long lists of biomolecules and their associated statistical dependencies. Hence, the numeric output of ARACNE, which contains MI values for all pairwise associations, was imported into Cytoscape for visualization, a conventional way of analysing interaction networks. Briefly, we reconstruct the network neighbourhood of each of the bioenergetic 'seed' genes listed in **Table S2** (i.e. all variables directly connected to them). The neighbouring variables can either be genes expressed in muscle and/or physiological variables. **Figure 3** summaries key regulatory associations (based on MI) between this seed set of genes and their immediate neighbours.

4. Functional analysis of the network hubs

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

[Figure 3 to be inserted here]

5. Biological interpretation

The most important finding of the current analysis is that among the direct neighbours to each bioenergetic pathway, particularly the two oxidative ones, we noted a statistical over-representation of genes encoding histone deacetylase enzymes (i.e. HDAC and SIRT mRNAs). This observation is consistent with previous studies that have highlighted the importance of sirtuins in regulating metabolism (15, 22, 28). Further, the protein deacetylase SIRT3 that primarily is localized in the mitochondrial matrix was also significantly positively correlated to both arterial oxygen tension (PaO_2) and oxygen uptake (VO_{2max}). In support of deacetylation being an important control point, it was recently shown that Sirt3 knockout mice exhibit decreased oxygen consumption, thus affecting cellular respiration (28). Hence, besides the obvious oxygen-driven effect on aerobic pathways (as indirect measures of oxygen availability such as VO_{2max} are linked to key genes in oxidative phosphorylation), the present network-based Systems Biology approach points to tissue hypoxia as being a potential important player in modifying expression of deacetylase modifying enzymes in severe COPD patients with a muscle wasting phenotype. Our Systems Biology approach also negatively links protein carbonylation (an established proxy measure for oxidative stress; (58)) to Complex 1 and 3 in the electron transport chain (**Figure 3**, bottom left part). The validity of such an association is further strengthened via functional enrichment analysis using DAVID, as a significant fraction of direct neighbouring genes to protein carbonylation is statistically associated to gene ontology (GO) terms representing cellular respiration.

If we then focus on the genes in the glycolytic pathway (**Figure 3**, top right part), a high proportion of pro-inflammatory mediators/receptors (e.g. IL1B, IL1R1 and TNFRSF21) are among the direct neighbours, as indicated by the enrichment of the ‘inflammatory response’ GO term (**Figure 3**, box A). Hence, hypoxia is pro-inflammatory, as seen by more traditional observation methods (20).

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

Step 2. Gene expression dynamics in response to tissue hypoxia

Animal models are commonly used for studying the in vivo effects of hypoxia, for ethical reasons, where severe or prolonged hypoxaemia is induced and invasive samples are required to explore mechanisms. Importantly, hindlimb skeletal muscles have been reported to alter metabolic phenotype and reduce fibre size in response to prolonged hypoxic stress in mice (53, 63), highlighting their potential relevance as a pre-clinical model of muscle wasting in COPD patients. In order to experimentally test the hypothesis derived from the clinical COPD network presented in **Figure 3**, we therefore exposed adult male C57/Bl6 mice to chronic systemic hypoxia for up to 2 weeks, in order to simulate levels of hypoxaemia reported in COPD patients with advanced respiratory insufficiency. To capture the temporal effect of reduced oxygen tension on gene regulation, we sampled and gene profiled the soleus muscle (n=4) at 3 different time-points (day 3, 7 and 14) following initiation of the gradual hypoxic insult (i.e. the O_2 level was gradually lowered to 10% over the first week and kept stable during the second week) (**Figure 2**, bottom part).

First, a high-level representation of the temporal transcriptional changes was performed using a variable reduction technique called principal component analysis (PCA) (**Figure 4B**). When plotting replicates of two variables against each other, it is relatively easy to see which is a better discriminating factor; visual inspection becomes increasingly difficult as the number of variables increase, hence the need for PCA. In essence, this method aims at ‘tilting’ the axes through the multidimensional data space, such that the first principal component accounts for as much of the variation in the original dataset as possible (the assumption is that the most important dynamics in the dataset are the ones with the largest variation). Our PCA revealed that the early dynamics of hypoxia is captured by the first principal component whereas the 2nd most important principal component (in terms of variance captured) separated the later time-points. Further, functional enrichment analysis of the differentially expressed genes (ANOVA, $P<0.05$) using DAVID (**Figure 4A**), highlighted several important pathways/ontologies. Most striking was the enrichment of protein catabolic process and ubiquitin-mediated proteolysis among genes up-regulated at day 7 and 14, clearly suggestive of a transcriptionally regulated muscle wasting phenotype driven by the experimentally induced hypoxaemic state.

State space models (SSMs) are a class of probabilistic graphical models (Koller and Friedman, 2009). SSM provides a general framework for analyzing deterministic and stochastic dynamical systems that can be measured/observed through a stochastic process. The SSM framework has been successfully used for the analysis of gene expression data (23, 51). In its simpler application the model formalises the effect of hidden, unmeasurable factors in specifying observed gene expression changes over time. The inclusion of these hidden factors is important since we cannot hope to measure all possible factors contributing to genetic regulatory interactions (e.g. levels of regulatory proteins as well as effects of mRNA and protein degradation).

The next step was to apply state-space modelling to reverse engineer transcriptional network modules (i.e. representing discrete temporal dimensions) from our replicated murine time-course dataset. Such module-based reduction in complexity allows analysis of hundreds or even thousands of genes, as those with a similar temporal expression profile are aggregated into a transcriptional module. To allow construction of a near genome-level model, we took advantage of a newer approach that incorporates this concept of modularization (23).

A state space model can reconstruct the topology of a network representing the systems dynamics, despite a relatively small number of time-points, by using biological replicates for each time-point (23). In order to reduce complexity, variables that do not change significantly are excluded from the modelling process. In this case study, genes deemed to be significant by ANOVA at a 1% significance level, as well as all hub genes listed in Table S3, were included (931 variables in total). The hub genes were chosen to represent the different components in our interpretative model derived from the clinical COPD dataset (**Figure 3**). Finally, the experimentally set oxygen level was used as an independent variable.

[Figure 5 to be inserted here]

Based on unsupervised clustering using HOPACH within the software programming environment R (35), we identified 8 distinct gene clusters with similar expression profiles. Hence, to model the effect of hypoxaemia on the skeletal muscle transcriptome the hidden state dimension was set to 4, as each inferred module contains both a positive (+) and a negative (-) component.

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

[Figure 6 to be inserted here]

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

CONCLUSIONS

The aim of this brief review is to provide an intuitive overview on data-driven 'learning' of biological pathways, linking molecular and physiological readouts. We used a case study to make it easier for experimental biologists to see the potential of computational biology to provide interpretative models of complex patterns, and stress that the identification of general properties of a system from a genome wide analysis of a molecular state of a system is a very powerful approach.

The ability to generate omics data with relatively accessible technologies offer an unprecedented opportunity to study how genetic information is used to control complex biological processes and their interaction. Until now we have only been able to understand a fraction of that complexity. The computational methods described in this review are designed to support this effort in the measure that they help isolate from these large datasets molecular signatures that correlate to phenotypic outcome.

With the help of computational biology, we are therefore able to develop hypothesis, which can be experimentally validated. In this context data-driven biology is not in contraposition with hypothesis-driven research. Instead it is a tool that support hypothesis generation in the event that the data is too complex to be interpreted solely using common sense. This approach is well developed in other areas of science, such as cancer biology where there is a vast literature showing that important hypothesis can be generated from modelling of these large datasets (12, 62).

In this manuscript, we demonstrate the development of an integrative workflow that incorporates measurements from different levels of cellular and molecular organization using a case study representing muscle wasting in COPD. The outline provides an exemplar where individual steps can be modified according to the type of data at hand and addition data types added. For example, in contrast to established gene expression microarrays, techniques for proteomics and especially metabolomics are still under development. Once it is possible to measure the whole proteome and metabolome of a sample, systems identification pipelines will clearly benefit from these omics techniques.

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

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

Figure legends

Figure 1: Schematic representation of the process involved in modelling a biological system by integrating knowledge from various sources, and complex multi-level datasets. The process can be conceptually subdivided into four distinct yet interconnected approaches (A-D). The experimental data used can either be novel multivariate data generated in your own (wet) laboratory or taken from a public repository. These may then be used to identify predictive biomarkers, i.e. variables that are predictive of a defined outcome (e.g. response to exercise training), and also to inform development of important networks that infer such outcomes; experimental data and other source of biological knowledge may also be useful in refining these representation of complex interactions. Such networks may in turn aid biomarker discovery, but are an essential precursor to computational models that are able to explore underlying molecular mechanisms; again, knowledge of specific biological issues may help in their refinement. Finally, incorporation of these models into larger scale analyses offer the potential for in silico experimentation, whereby e.g. the effect of different therapeutic interventions on disease outcome may be tested.

Figure 2: Schematic representation of the analysis strategy used in the case study, highlighting how the inferred static multi-scale network from the clinical COPD cohort (Fig 2A-C) can be bridged to the inference of a dynamical network representing the temporal progression of events following an experimental challenge (hypoxic exposure) in a murine animal model (Fig 2D-G). Having identified a clinical condition with known outcome (exercise intolerance in patients with respiratory disease), we could target unknown mechanisms by focussing on one likely source of functional limitation (skeletal muscle dysfunction ± central limitation on O_2 supply),

and generate data characterising the phenotype. Both genomic and physiological readouts were used to construct a network of inferred interactions, which was then interrogated to identify statistically robust linkages among broad biological functions. While very useful in providing a list of useful biomarkers, there remains a potential limitation with single point associations. The dynamic nature of relationships is captured by repeated measures across a suitable time scale (which will vary for different molecular, physiological and structural responses) using an animal model of respiratory distress, where the transcriptome-based model demonstrated the central importance of oxygen in the response.

Figure 3: Graphical representation highlighting putative regulatory associations (significant correlation between two factors is shown as a dotted line) that likely represent robust interactions, based on high mutual information values. The focus is on central metabolism pathways (i.e. glycolysis, TCA and OXPHOS, respectively) and their immediate neighbours. The grey boxes define functional enrichment of the different bioenergetic compartments based on direct neighbours. Individual genes of relevance are grouped into modules with others of related function, as are physiological readouts that may be treated in a similar manner for statistical analysis. C1-5: the different complexes in the electron transport chain. The value of such an approach is in providing a detailed overview of a complex interaction network, reducing the huge number of potential factors into groups of defined function, and offering a limited number of candidates whose utility as biomarkers or therapeutic targets may be experimentally verified.

Figure 4: High-level representation of temporal transcriptional changes in the murine model of hypoxia. A) Graphical representation of the pre-clinical experimental design. The oxygen level was gradually decreased from 21% to 10% during the first week and mice were housed for another week at this oxygen concentration. B) Principal component plot highlighting the transcriptional dynamics caused by the hypoxic challenge. C) Hierarchical clustering using mRNA expression levels of genes modulated by hypoxia ($P<0.05$). Each row represents a transcript and each column represents a sample. Red and green colours indicate expression levels above and below the median value of the distribution of signal, respectively. Using solid yellow lines we have subdivided genes into overall trends in order to help the reader. Enriched functional terms within these are listed next to the heatmap.

Figure 5: The hierarchical dynamic state-space model identified 4 modules (x-axes define length of hypoxic exposure), each characterised by two separate transcriptional profiles: plus and minus, representing up- and down-regulation, respectively. The hierarchical position of the modules represents the estimated temporal structure of the network. Functionally enriched GO terms (regular text) as well as key genes (italics) are identified next to the relevant module. Blue arrows represent temporal repression whereas red arrows represent temporal induction. The numeric value next to each arrow represents the estimated coefficient.

Figure 6: A higher resolution representation of Figure 5, highlighting the most significant gene interactions between components in the four inferred modules. Lines represent factor interactions based on mutual information (blue represents temporal repression, red represents temporal induction). Genes are colour coded for broad

functional categories (red=cytokines; blue=epigenetic modifiers; green=aerobic metabolism; purple=muscle differentiation; yellow=cell-interaction).

References:

1. **Alderson D**, **Doyle JC**, **Li L**, **Willinger W**. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. Internet Math 2: 431–523, 2005.

2. **Ashburner M**, **Ball CA**, **Blake JA**, **Botstein D**, **Butler H**, **Cherry JM**, **Davis AP**, **Dolinski K**, **Dwight SS**, **Eppig JT**, **Harris MA**, **Hill DP**, **Issel-Tarver L**, **Kasarskis A**, **Lewis S**, **Matese JC**, **Richardson JE**, **Ringwald M**, **Rubin GM**, **Sherlock G**. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–9, 2000.

3. **Bansal M**, **Della Gatta G**, **di Bernardo D**. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics 22: 815–22, 2006.

4. **Barabasi A**, **Albert R**. Emergence of scaling in random networks [Online]. Science 286: 509–12, 1999. http://www.ncbi.nlm.nih.gov/pubmed/10521342 [26 Aug. 2015].

5. **Barrett T**, **Suzek TO**, **Troup DB**, **Wilhite SE**, **Ngau W-C**, **Ledoux P**, **Rudnev D**, **Lash AE**, **Fujibuchi W**, **Edgar R**. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 33: D562–6, 2005.

6. **Batagelj V**, **Mrvar A**. Pajek - Program for large network analysis. Connections 21: 47–57, 1998.

7. **Brazma A**, **Parkinson H**, **Sarkans U**, **Shojatalab M**, **Vilo J**, **Abeygunawardena N**, **Holloway E**, **Kapushesky M**, **Kemmeren P**, **Lara GG**, **Oezcimen A**, **Rocca-Serra P**, **Sansone S-A**. ArrayExpress--a public repository for microarray gene expression data at the EBI. [Online]. Nucleic Acids Res 31: 68–71, 2003. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165538&tool=pmcentrez&rendertype=abstract [18 Jun. 2013].

8. **Butte AJ**, **Kohane IS**. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. [Online]. Pac. Symp. Biocomput. http://www.ncbi.nlm.nih.gov/pubmed/10902190 [23 Jul. 2013].

9. **Cassese A**, **Guindani M**, **Tadesse MG**, **Falciani F**, **Vannucci M**. A HIERARCHICAL BAYESIAN MODEL FOR INFERENCE OF COPY NUMBER VARIANTS AND THEIR ASSOCIATION TO GENE EXPRESSION. Ann Appl Stat 8: 148–175, 2014.

10. **Davidsen PK**, **Herbert JM**, **Antczak P**, **Clarke K**, **Ferrer E**, **Peinado VI**, **Gonzalez C**, **Roca J**, **Egginton S**, **Falciani F**. A systems biology approach reveals a link between systemic cytokines and skeletal muscle energy metabolism in a rodent smoking model and human COPD. Genome Med. .

11. **Dennis G**, **Sherman BT**, **Hosack DA**, **Yang J**, **Gao W**, **Lane HC**, **Lempicki RA**. DAVID: Database for Annotation, Visualization, and Integrated Discovery. [Online]. Genome Biol 4: P3, 2003. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3720094&tool=pmcentrez&rendertype=abstract [5 Jun. 2014].

12. **Du W**, **Elemento O**. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene 34: 3215–25, 2015.

13. **Faith JJ**, **Hayete B**, **Thaden JT**, **Mogno I**, **Wierzbowski J**, **Cottarel G**, **Kasif S**, **Collins JJ**, **Gardner TS**. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5: e8, 2007.

14. **Falati S**, **Gross P**, **Merrill-Skoloff G**, **Furie BC**, **Furie B**. Real-time in vivo imaging of platelets, tissue factor and fibrin during arterial thrombus formation in the mouse. Nat Med 8: 1175–1180, 2002.

15. **Finkel T**, **Deng C-X**, **Mostoslavsky R**. Recent progress in the biology and physiology of sirtuins. Nature 460: 587–91, 2009.

16. **Gavaghan D**, **Garny A**, **Maini PK**, **Kohl P**. Mathematical models in physiology. Philos Trans A Math Phys Eng Sci 364: 1099–106, 2006.

17. **Gehlenborg N**, **O’Donoghue SI**, **Baliga NS**, **Goesmann A**, **Hibbs MA**, **Kitano H**, **Kohlbacher O**, **Neuweger H**, **Schneider R**, **Tenenbaum D**, **Gavin A-C**. Visualization of omics data for systems biology. Nat Methods 7: S56–68, 2010.

18. **Goldovsky L**, **Cases I**, **Enright AJ**, **Ouzounis CA**. BioLayout(Java): versatile network visualisation of structural and functional relationships. [Online]. Appl Bioinformatics 4: 71–4, 2005. http://www.ncbi.nlm.nih.gov/pubmed/16000016 [15 Jul. 2014].

19. **Gomez-Cabrero D**, **Compte A**, **Tegner J**. Workflow for generating competing hypothesis from models with parameter uncertainty. Interface Focus 1: 438–49, 2011.

20. **Gonzalez NC**, **Wood JG**. Alveolar hypoxia-induced systemic inflammation: what low PO(2) does and does not do. Adv Exp Med Biol 662: 27–32, 2010.

21. **Gupta R**, **Stincone A**, **Antczak P**, **Durant S**, **Bicknell R**, **Bikfalvi A**, **Falciani F**. A computational framework for gene regulatory network inference that combines multiple methods and datasets. BMC Syst Biol 5: 52, 2011.

22. **He W**, **Newman JC**, **Wang MZ**, **Ho L**, **Verdin E**. Mitochondrial sirtuins: regulators of protein acylation and metabolism. Trends Endocrinol Metab 23: 467–76, 2012.

23. **Hirose O**, **Yoshida R**, **Imoto S**, **Yamaguchi R**, **Higuchi T**, **Charnock-Jones DS**, **Print C**, **Miyano S**. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. Bioinformatics 24: 932–42, 2008.

24. **Holloway K V**, **O’Gorman M**, **Woods P**, **Morton JP**, **Evans L**, **Cable NT**, **Goldspink DF**, **Burniston JG**. Proteomic investigation of changes in human vastus lateralis muscle in response to interval-exercise training. Proteomics 9: 5155–74, 2009.

25. **Huang DW**, **Sherman BT**, **Lempicki RA**. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13, 2009.

26. **HUXLEY AF**. Muscle structure and theories of contraction. [Online]. Prog Biophys Biophys Chem 7: 255–318, 1957. http://www.ncbi.nlm.nih.gov/pubmed/13485191 [15 Jul. 2014].

27. **Ideker T**, **Ozier O**, **Schwikowski B**, **Siegel AF**. Discovering regulatory and signalling circuits in molecular interaction networks. [Online]. Bioinformatics 18 Suppl 1: S233–40, 2002. http://www.ncbi.nlm.nih.gov/pubmed/12169552 [3 Jun. 2014].

28. **Jing E**, **Emanuelli B**, **Hirschey MD**, **Boucher J**, **Lee KY**, **Lombard D**, **Verdin EM**, **Kahn CR**. Sirtuin-3 (Sirt3) regulates skeletal muscle metabolism and insulin signaling via altered mitochondrial oxidation and reactive oxygen species production. Proc Natl Acad Sci U S A 108: 14608–13, 2011.

29. **Jones P**, **Côté RG**, **Martens L**, **Quinn AF**, **Taylor CF**, **Derache W**, **Hermjakob H**, **Apweiler R**. PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic Acids Res 34: D659–63,

2006.

30. **Joyner MJ**, **Pedersen BK**. Ten questions about systems biology. J Physiol 589: 1017–30, 2011.

31. **Keller P**, **Vollaard NBJ**, **Gustafsson T**, **Gallagher IJ**, **Sundberg CJ**, **Rankinen T**, **Britton SL**, **Bouchard C**, **Koch LG**, **Timmons JA**. A transcriptional map of the impact of endurance exercise training on skeletal muscle phenotype. J Appl Physiol 110: 46–59, 2011.

32. **Kohl P**, **Noble D**. Systems biology and the virtual physiological human. Mol Syst Biol 5: 292, 2009.

33. **Krogh A**. The number and distribution of capillaries in muscles with calculations of the oxygen pressure head necessary for supplying the tissue. [Online]. J Physiol 52: 409–15, 1919. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1402716&tool=pmcentrez&rendertype=abstract [4 Jul. 2014].

34. **Kupershmidt I**, **Su QJ**, **Grewal A**, **Sundaresh S**, **Halperin I**, **Flynn J**, **Shekar M**, **Wang H**, **Park J**, **Cui W**, **Wall GD**, **Wisotzkey R**, **Alag S**, **Akhtari S**, **Ronaghi M**. Ontology-based meta-analysis of global collections of high-throughput public data. PLoS One 5: e13066, 2010.

35. **van der Laan MJ**, **Pollard KS**. Hybrid clustering of gene expression data with visualization and the bootstrap. J Stat Plan Inference 117: 275–303, 2003.

36. **Maere S**, **Heymans K**, **Kuiper M**. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21: 3448–9, 2005.

37. **Mah N**. A comparison of oligonucleotide and cDNA-based microarray systems. Physiol Genomics 16: 361–370, 2004.

38. **Margolin AA**, **Nemenman I**, **Basso K**, **Wiggins C**, **Stolovitzky G**, **Dalla Favera R**, **Califano A**. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. [Online]. BMC Bioinformatics 7 Suppl 1: S7, 2006. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1810318&tool=pmcentrez&rendertype=abstract [19 Jul. 2012].

39. **Mattson DL**. Functional Genomics. In: Integrative Physiology in the Proteomics and Post-Genomics Age, edited by Walz W. Humana Press Inc.,

2005, p. 7–26.

40. **Merico D**, **Gfeller D**, **Bader GD**. How to visually interpret biological data using networks. Nat Biotechnol 27: 921–924, 2009.

41. **Meyer PE**, **Kontos K**, **Lafitte F**, **Bontempi G**. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J. Bioinform. Syst. Biol. (January 2007). doi: 10.1155/2007/79879.

42. **Neapolitan RE**. Learning Bayesian Networks. New Jersey: Pearson Prentice Hall, 2004.

43. **Noble D**. The Music of Life: Biology beyond genes. Oxford University Press, 2008.

44. **Noble D**. Computational models of the heart and their use in assessing the actions of drugs. [Online]. J Pharmacol Sci 107: 107–17, 2008. http://www.ncbi.nlm.nih.gov/pubmed/18566519 [7 Jul. 2014].

45. **Ogata H**, **Goto S**, **Sato K**, **Fujibuchi W**, **Bono H**, **Kanehisa M**. KEGG: Kyoto Encyclopedia of Genes and Genomes. [Online]. Nucleic Acids Res 27: 29–34, 1999. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148090&tool=pmcentrez&rendertype=abstract [17 Jun. 2013].

46. **Opgen-Rhein R**, **Strimmer K**. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. BMC Bioinformatics 8 Suppl 2: S3, 2007.

47. **Ortega F**, **Sameith K**, **Turan N**, **Compton R**, **Trevino V**, **Vannucci M**, **Falciani F**. Models and computational strategies linking physiological response to molecular networks from large-scale data. Philos Trans A Math Phys Eng Sci 366: 3067–89, 2008.

48. **Perrin B-E**, **Ralaivola L**, **Mazurie A**, **Bottani S**, **Mallet J**, **d'Alché-Buc F**. Gene networks inference using dynamic Bayesian networks. [Online]. Bioinformatics 19 Suppl 2: ii138–48, 2003. http://www.ncbi.nlm.nih.gov/pubmed/14534183 [2 Nov. 2014].

49. **Poultney CS**, **Greenfield A**, **Bonneau R**. Integrated inference and analysis of regulatory networks from multi-level measurements. Methods Cell Biol 110: 19–56, 2012.

50. **Rabinovich RA**, **Bastos R**, **Ardite E**, **Llinàs L**, **Orozco-Levi M**, **Gea J**, **Vilaró J**, **Barberà JA**, **Rodríguez-Roisin R**, **Fernández-Checa JC**, **Roca J**. Mitochondrial dysfunction in COPD patients with low body mass index. Eur Respir J 29: 643–50, 2007.

51. **Rangel C**, **Angus J**, **Ghahramani Z**, **Lioumi M**, **Sotheran E**, **Gaiba A**, **Wild DL**, **Falciani F**. Modeling T-cell activation using gene expression profiling and state-space models. Bioinformatics 20: 1361–1372, 2004.

52. **Ravasz E**, **Somera AL**, **Mongru DA**, **Oltvai ZN**, **Barabási AL**. Hierarchical organization of modularity in metabolic networks. Science 297: 1551–5, 2002.

53. **Reinke C**, **Bevans-Fonti S**, **Drager LF**, **Shin M-K**, **Polotsky VY**. Effects of different acute hypoxic regimens on tissue oxygen profiles and metabolic outcomes. J Appl Physiol 111: 881–90, 2011.

54. **Segal E**, **Shapira M**, **Regev A**, **Pe'er D**, **Botstein D**, **Koller D**, **Friedman N**. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166–76, 2003.

55. **Shannon P**, **Markiel A**, **Ozier O**, **Baliga NS**, **Wang JT**, **Ramage D**, **Amin N**, **Schwikowski B**, **Ideker T**. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–504, 2003.

56. **De Smet R**, **Marchal K**. Advantages and limitations of current network inference methods. Nat Rev Microbiol 8: 717–29, 2010.

57. **Su G**, **Kuchinsky A**, **Morris JH**, **States DJ**, **Meng F**. GLay: community structure analysis of biological networks. Bioinformatics 26: 3135–7, 2010.

58. **Suzuki YJ**, **Carini M**, **Butterfield DA**. Protein carbonylation. Antioxid Redox Signal 12: 323–5, 2010.

59. **Timmons JA**, **Knudsen S**, **Rankinen T**, **Koch LG**, **Sarzynski M**, **Jensen T**, **Keller P**, **Scheele C**, **Vollaard NBJ**, **Nielsen S**, **Akerström T**, **MacDougald OA**, **Jansson E**, **Greenhaff PL**, **Tarnopolsky MA**, **van Loon LJC**, **Pedersen BK**, **Sundberg CJ**, **Wahlestedt C**, **Britton SL**, **Bouchard C**. Using molecular classification to predict gains in maximal aerobic capacity following endurance exercise training in humans. [Online]. J Appl Physiol 108: 1487–96, 2010. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2886694&tool=pmcentrez&rendertype=abstract [13 Jul. 2012].

60. **Trevino V**, **Falciani F**. GALGO: an R package for multivariate variable selection using genetic algorithms. Bioinformatics 22: 1154–6, 2006.

61. **Turan N**, **Kalko S**, **Stincone A**, **Clarke K**, **Sabah A**, **Howlett K**, **Curnow SJ**, **Rodriguez DA**, **Cascante M**, **O'Neill L**, **Egginton S**, **Roca J**, **Falciani F**. A systems biology approach identifies molecular networks defining skeletal muscle abnormalities in chronic obstructive pulmonary disease. PLoS Comput Biol 7: e1002129, 2011.

62. **Werner HMJ**, **Mills GB**, **Ram PT**. Cancer Systems Biology: a peek into the future of patient care? Nat Rev Clin Oncol 11: 167–176, 2014.

63. **Willmann G**. Transcriptional Regulation after Chronic Hypoxia Exposure in Skeletal Muscle. University of Cologne: 2013.

64. **Yu J**, **Smith VA**, **Wang PP**, **Hartemink AJ**, **Jarvis ED**. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20: 3594–603, 2004.