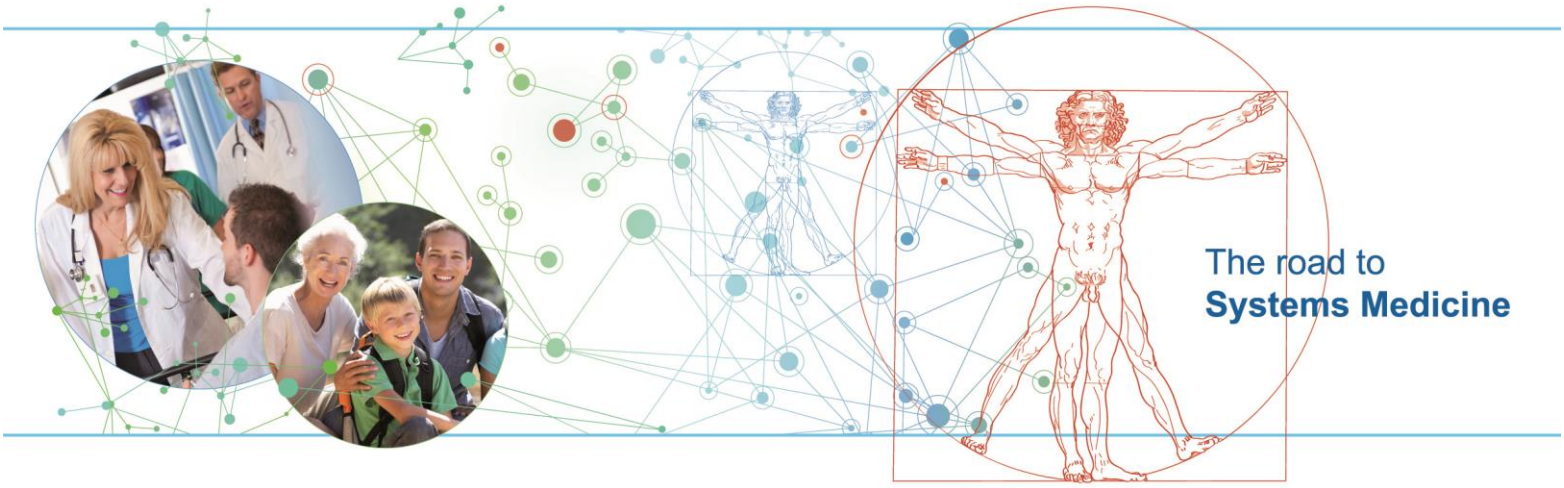
Coordinating Action Systems Medicine
Implementation of Systems Medicine across Europe



Workshops on genomic models of complex disease genes

CASyM report

November 2014

IMPRINT

Publisher

CASyM administrative office
Project Management Jülich, Forschungszentrum Jülich GmbH
m.kirschner@fz-juelich.de

Authors

Mikael Benson, Mika Gustafsson, Colm Nestor, Huan Zhang *(Linköping University, Sweden)*

Date

November 2014 (updated: March 2016)

Contact information

Mikael Benson
Linköping University, Sweden
mikael.benson@liu.se

TABLE OF CONTENT

WORKSHOPS ON GENOMIC MODELS OF COMPLEX DISEASE GENES

Summary

As highlighted by the CASyM roadmap a key problem for translation of systems medical research to the clinic, is the availability of the enormous amount of disease relevant high-throughput data is available in public repositories. This is clearly relevant for clinical research but key challenge lie in systematically identifying optimal datasets, availability in accessible formats, which can be translated to clinical research, for example to find biomarkers for personalised medicine. Addressing this challenge requires multi-disciplinary collaborations between experts in omics, bioinformatics, functional and clinical research. The aims of task 3.4 are to introduce principles how to access and analyse clinically relevant omics data in a user-friendly way that is easy to translate to clinical research. To reach these aims we have introduced module-based genomic models of complex disease genes and their regulators. The models have been presented at several workshops and will serve as a basis for courses introducing systems medicine to clinicians in WP2, as well as for clinicians interested in applying systems medicine to their research on specific diseases. The models have been assembled based on published reports and data in the public domain, including microarray data available in public repositories such as GEO, http://www.ncbi.nlm.nih.gov/geo/. An important principle behind the models is to use network-based concepts to organise disease associated genes. Briefly, disease-associated genes from each disease tend to co-localise on the human protein-protein interaction network and form modules of functionally related genes. Those modules can be used as a reference for clinical researchers to find disease-mechanisms, as well as diagnostic markers and therapeutic targets.

Background

Despite impressive advances during the last century, modern health care is faced with enormous challenges. One problem is that currently available drugs show highly variable clinical efficacy, which results not only in suffering, but also contributes to increasing costs. The annual cost of ineffective prescriptions currently on the market is at an estimated $350 billion [1]. Variable efficacy also adds to the huge costs associated with drug discovery, development and clinical trials (on average one billion dollars per drug), which further impacts the financing of health care. These problems reflect the complexity of common diseases, which can involve altered interactions between thousands of genes. Because of the large number of genes and their interconnection, it is very difficult to gain functional understanding of disease mechanisms by detailed studies of individual genes.

This problem of complexity is compounded by disease heterogeneity: patients with similar clinical manifestations may have different underlying disease mechanisms. Asthma is an example of such a disease; it can be caused by infection, allergens or other environmental factors, which give rise to different inflammatory responses. Variations in response may underlie the observation that between 10 and 20% of patients do not respond to one of the most common asthma drugs, corticosteroids. This variation, however, can potentially be exploited to find novel drugs for non-responders in asthma, allergy and other diseases, as well as to identify patients that require such drugs.

Systems medicine is an emerging discipline that aims to address the problem that a disease is rarely caused by malfunction of one individual gene product, but instead depends on multiple gene products that interact in a

complex network. High throughput technologies allow analysis of all human disease associated genes and gene products. Data from such technologies are available in public databases, which allow the construction of genomic models of common diseases. We have introduced systems medical principles and presented such models in various workshops and conferences to basic and clinical researchers, as well as medical students. This has required collaborative efforts from a team of multi-disciplinary researchers with expertise in omics, bioinformatics, functional and clinical research. The resulting principles, and the clinical implications have also been presented in review articles by Zhang et al. and Gustafsson et al. in 2014, or Benson et al in 2015 [2-4], as well as in mass media, including opinion articles, TV and radio interviews during 2014-2016. Below, we summarize the contents of these introductory talks. We explain how and why systems medicine, and specifically network approaches, can be used to assist clinical decision making and to identify underlying disease mechanisms. We focus on the use of disease modules to uncover pathogenic mechanisms and describe how these can be extended into multilayer networks.

Disease relevant data and tools in the public domain

One of the largest sources is MEDLINE. Biomedical researchers routinely search MEDLINE to find disease relevant genes and information about their mechanisms. However, manual searches are complicated by the large number of genes and articles. The human genome contains some 20,000 genes that are highly interconnected. MEDLINE contains some 20 million abstracts. One way to address this complexity is computational mining of abstracts in MEDLINE. Briefly, the principle is that if two genes are mentioned in the same abstract they are likely to be functionally related. Statistical methods can be used to sort out significant co-citations, and link such genes into networks, in which links represent putative interactions [5]. The authors made the resulting network model freely available in the public domain, together with search tools, so that it is possible to search for associations between diseases and genes:

- http://www.coremine.com/medical/#search

Another large and important source of disease-relevant data is the Online Mendelian Inheritance in Man (OMIM). This is a manually curated, online catalogue of human genes and genetic disorders, which is continuously updated:

- http://www.omim.org

The NHGRI GWAS catalogue contains lists of disease associated genes from genome wide association studies (GWAS):

- http://www.genome.gov/gwastudies/

The Gene Expression Omnibus (GEO) contains hundreds of thousands of mRNA microarrays experiments but also several other forms of high throughput experiments, such as microRNAs and DNA methylation:

- http://www.ncbi.nlm.nih.gov/geo/

The human protein atlas contains information about a large number of proteins, their functions and cellular or tissue distribution, as well as antibodies to detect them:

- http://www.proteinatlas.org

In order to organize disease-associated genes into networks, models of human protein-protein interaction networks are needed. Several such models are available in the public domain, for example HPRD, Reactome, Intact and HI2 in the pleiotropic module that was also present in the largest connected component. Another organizing principle is to search for pathway enrichment, for example among GWAS genes. Several free or commercial pathway databases are currently available, namely Kyoto Encyclopedia of Genes and Genomes (KEGG), Ingenuity Pathway Analysis, and Gene Ontology (GO). The analysis of mouse knockout phenotypes was performed by downloading phenotypic information from [36] as of 31 January 2013.

Drug target databases allow systematic searches poor genes targeted by specific drugs, for example in cancer:

- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3349233/?tool=pubmed

Mouse knockout databases allow identification of putative mammalian phenotypes resulting from specific gene aberrations:

- ftp://ftp.informatics.jax.org/pub/reports/index.html#pheno
- http://www.genoway.com/services/eucomm/eucomm-conditional-knockouts.htm?utm_source=google&utm_medium=cpc&utm_campaign=europe

Networks and modules to organize disease associated genes

Networks provide graphical representations of complex systems. In the context of cellular networks, molecules such as genes and proteins are represented as nodes, and the interactions among them as links. In a landmark article in 1999 by Barabasi et al. [6], it was shown that networks in technological, social, and biological systems have common designs that are governed by simple and quantifiable organising principles. Key findings were that a fraction of the nodes served as hubs with multiple links, whereas the vast majority of nodes had few links. The hubs often had large individual effects, in contrast to the nodes with few links. The hubs contributed to the small world property of networks: all nodes in a network are generally connected by a limited number of links. Another important characteristic is that functionally related nodes tended to be highly interconnected and co-localize in networks, thereby forming modules.

Genomic diseases models

In the context of disease, disease-associated genes identified by omics studies can be computationally mapped on to models of the human protein-protein interaction (PPI) network. In other words, each disease-associated gene is mapped on to its matching protein product. The resulting maps have characteristics that are similar to those found in other types of networks. One of the most important characteristics is that functionally related genes tend to co-localize and form disease modules.

Disease modules can help to organise and prioritise disease-associated genes identified by high-throughput analyses (**Figure 1**), as well as to provide an overview of disease-mechanisms by performing pathway analyses. Disease modules can also help to identify novel disease genes, biomarkers or therapeutic targets. In 2007, Pujana *et al*. [7] described a module relevant to breast cancer, and identified a novel candidate gene, *HMMR*, that was validated by functional and genetic studies. In 2014, a module-based approach for drug discovery was described in rheumatoid arthritis based on a meta-analysis of GWAS of 100,000 subjects [8]. Analysis of disease modules exploits the general principles of networks such as alteration of hub genes being likely to have large effects, while alterations in the many genes with few links will likely correspond to small-effect genes. Thus, specific therapeutic targeting of a hub gene is more likely to be effective than a gene with few interactions. Indeed, genes targeted by drugs have more interactions than other genes, which increases the risk that a drug targeting a specific disease gene may have an off-target effect. An important observation is that nodes that are highly interconnected in a network are likely to be functionally related. Thus, novel candidate genes can be found among the interactors of known disease genes.

One recent example of a successful module-based approachwas based on the assumption that the genes in a module would be co-regulated by the same set of transcription factors (TFs) that regulate a known disease-gene *IL13* [9] (**Figure 2**). Twenty-five putative *IL13*-regulating TFs were knocked down using siRNA, of which seven were found to affect *IL13*. The knockdowns were repeated for these TFs, followed by mRNA microarrays to detect their downstream targets. This led to the identification of a module of highly interconnected genes. That module contained several genes of known relevance to allergy, such as *IFNG*, *IL12*, *IL4*, *IL5*, *IL13* and their receptors. It also contained novel candidate genes, including *S100A4,* which was validated as a diagnostic and therapeutic candidate by a combination of functional, mouse and clinical studies. A mouse knock-out model showed that *S100A4* had extensive phenotypic, cellular and humoral effects on allergic inflammation. The

therapeutic potential was demonstrated by treatment with a specific antibody, both in the mouse model and in cells from allergic patients.

The success of single module approaches in identifying candidate genes prompted researchers to extend it to multiple modules to link genomic, phenotypic and environmental variables together. Rapid development of high-throughput techniques has enabled global analyses of different network layers ranging from DNA to proteins, as well as metabolites and lipids. Similar to genes, the variables in each layer can be linked to each other. Consider, for example, one disease module formed by mRNAs and another from single nucleotide polymorphisms (SNPs). If an mRNA and a SNP in each module map to the same protein, they can be linked. This principle can be expanded to all proteins in the module and the overlap tested statistically. Another example is modules formed by genes and their regulators, such as transcription factors or microRNAs. Genes can be linked if they are regulated by the same microRNAs, and a double-layer module can then be formed by linking microRNAs that regulate the same gene. By combining different high-throughput analyses it is therefore possible to form multi-layer disease modules (MLDMs).

Multi-dimensional models can be used to form rejectable hypotheses of how genes, gene products and regulators interact with each other. For example, does a disease-associated SNP in a promoter region of a module gene change the expression of of that gene? Does a microRNA regulate its predicted target genes in a module? The clinical relevance of MLDMs lies in that they can provide a framework to identify optimal combinations of diagnostic markers from different layers, based on functional understanding of the pathogenic roles of those markers. For example, microRNAs and genetic variants were used to examine disease-associated variations in mRNA expression in gliomas, and to predict disease outcome.

An important aspect of MLDMs is that they can be linked to modules formed by other clinical data. For example, a link can be placed between a disease and a gene associated with that disease. Next, diseases that are associated with the same gene can be linked and form a human disease network. The same principle can be applied to the disease genes forming a disease gene network. Such networks are modular and can be linked, so that diseases can be associated with the underlying disease mechanisms. It is also possible to construct and link modules containing other relevant data, such as social and environmental factors, and use this for personalised medicine (**Figure 3**).

MLDMs might also be useful for tracking disease over multiple time points. Diseases are dynamic processes rather than static entities, and the underlying processes and time frames may range from hours in rapidly evolving cases, such as meningitis, to decades in cancer. Disease progression is perhaps best understood in cancer. For example, at a molecular level, a study of chronic lymphocytic leukemia revealed the development of substantial genetic heterogeneity of tumor cells from the same patients over time. Such developments were linked to disease deterioration and variable treatment response. Thus, understanding of module kinetics can be exploited for sequential treatment with different drugs. Ideally, this principle should be expanded so that all diseases are staged using MLDMs with omics and routine clinical data integrated. In the future, it may be possible to infer early MLDMs, before patients become symptomatic, allowing preventative medicine.

It is possible that personal MLDMs could become a cornerstone for health care, and could be used for the early diagnosis of changes in module function, based on functional understanding of why disease-causing nodes in the MLDMs change (such as due to a genetic variant). As the bioinformatics principles for analysing different forms of variables are largely the same, MLDMs could also include other forms of clinical information such as routine laboratory tests and medical imaging. The versatility and resolution of medical imaging is steadily increasing and is aiming to proivde functional understanding of observed structural changes in the human body.

In summary, MLDMs can potentially be used as templates to integrate and analyse multiple layers of disease-relevant information. Similar to the current diagnostic model discussed above, analyses can be based on functional understanding, but with higher resolution and the option for computational predictions. When the

underlying mechanisms are revealed, our view of various common diseases might alter, prompting reclassification of multiple diseases.

Figures

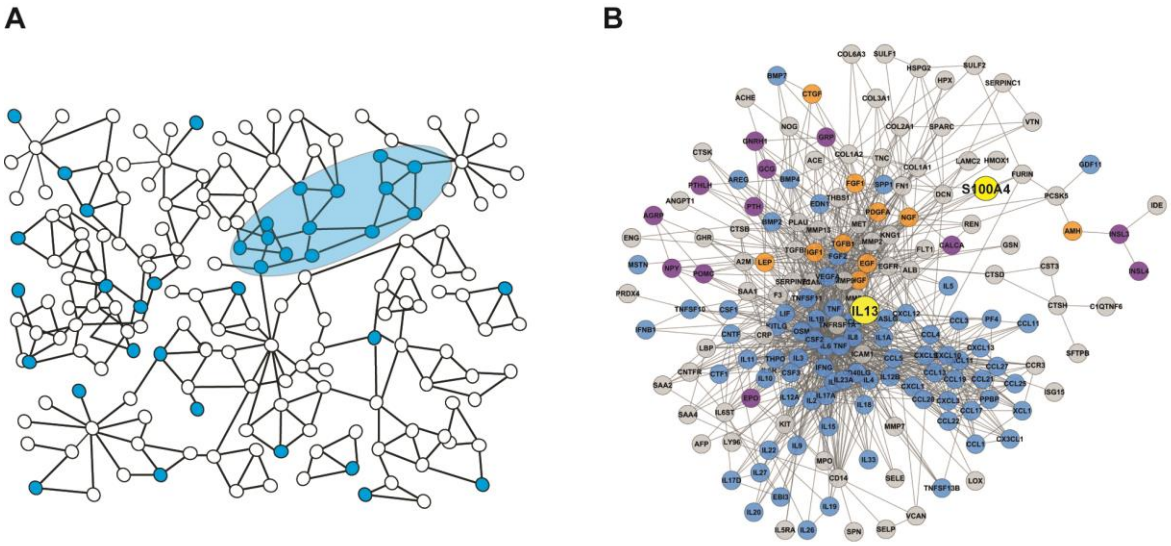
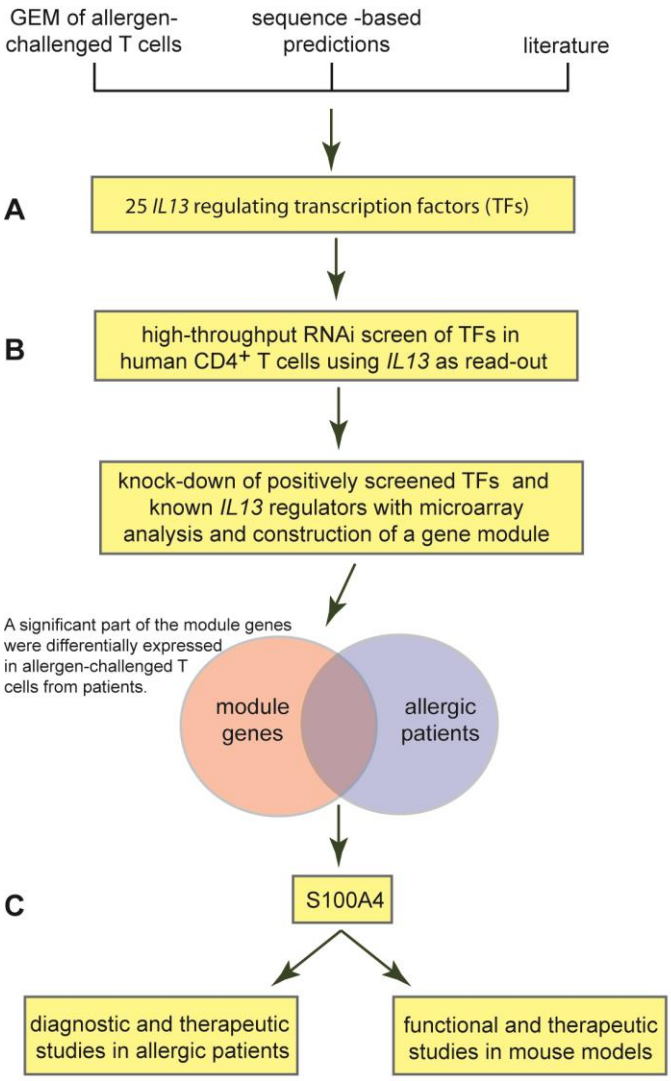
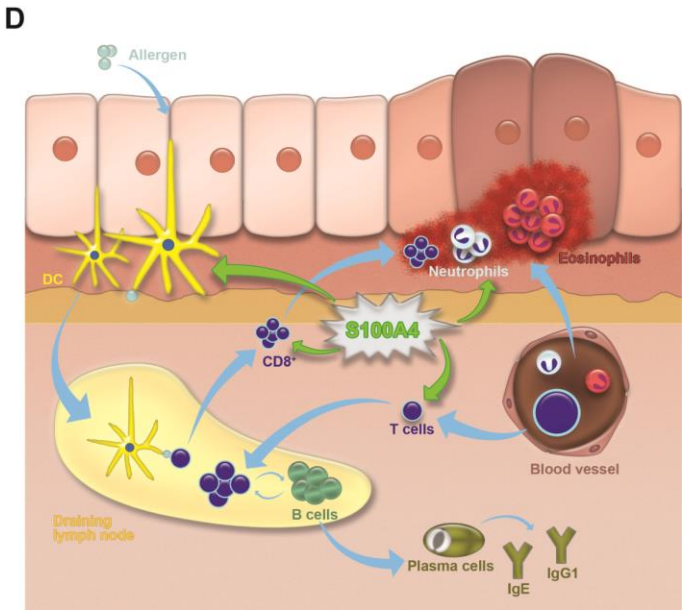


Figure 1. A disease module: A) Conceptual model of how disease-associated genes (blue nodes), identified by high-throughput analysis, tend to co-localise in the human protein-protein interaction network (white nodes), forming a module (blue oval). The genes in the module are assumed to be more important for the disease than extra-modular genes. **B**) An actual disease module from allergic patients, showing extracellular proteins that were putatively co-regulated with IL13. Blue nodes are associated with Cytokine Activity, purple nodes are associated with Hormone Activity, and orange nodes are associated with Growth Factor Activity according to Gene Ontology Molecular Function. Panel B is taken from [9].

Figure 2 (next page): A module-based approach to identify disease-relevant diagnostic and therapeutic candidate genes in allergy. A) Twenty-five putative *IL13*-regulating TFs were identified by combining data from mRNA microarrays, sequence-based predictions, and the literature. **B**) *IL13*-regulating TFs were validated by siRNA-mediated knockdown of the 25 TFs in human total $CD4^+$ T cells polarised toward T_H2 using *IL13* as a read-out. The target genes of the TFs were identified by combined siRNA knockdown of the positively screened TFs/known IL-13–regulating TFs from literature and microarray analyses. This resulted in a module of genes that was co-regulated with *IL13* in T_H2-polarised cells and significantly overlapped with differentially expressed genes from allergen-challenged T cells from allergic patients. For further validation experiments, the study focused on module genes that encoded secreted proteins and had not been previously associated with allergy. **C**) Functional, diagnostic, and therapeutic studies involving one of the module genes, *S100A4,* were performed in patients with SAR, allergic dermatitis, and a mouse model of allergy. **D**) Model of S100A4-induced disease mechanisms. Allergic inflammation requires the sensitisation of the immune system by allergens, resulting in the production of antigen-specific T cells. The interaction of dendritic cells (DCs) in the draining lymph node with T cells is a critical step that is dependent on S100A4. B-cell maturation as a result of T cell–B cell crosstalk (for example, the release of T_H2 cytokines by T cells) leads to the production of IgE and IgG1 by plasma cells. Cytokines and chemokines released by T cells stimulate the migration of circulating granulocytes (for example, neutrophils and eosinophils) to the inflammatory site (skin). Differentiation of naïve T cells into $CD8^+$ cytotoxic T cells will exacerbate the skin damage. Blue arrows indicate the flow of the allergic responses. Green arrows indicate the promotion of these processes by S100A4.


GEM of allergen-challenged T cells
sequence -based predictions
literature
A
25 IL13 regulating transcription factors (TFs)
B
high-throughput RNAi screen of TFs in human CD4+ T cells using IL13 as read-out
knock-down of positively screened TFs and known IL13 regulators with microarray analysis and construction of a gene module
A significant part of the module genes were differentially expressed in allergen-challenged T cells from patients.
module genes
allergic patients
C
S100A4
diagnostic and therapeutic studies in allergic patients
functional and therapeutic studies in mouse models


D
Allergen
DC
Neutrophils
Eosinophils
S100A4
CD8+
T cells
Blood vessel
B cells
Draining lymph node
Plasma cells
IgE
IgG1

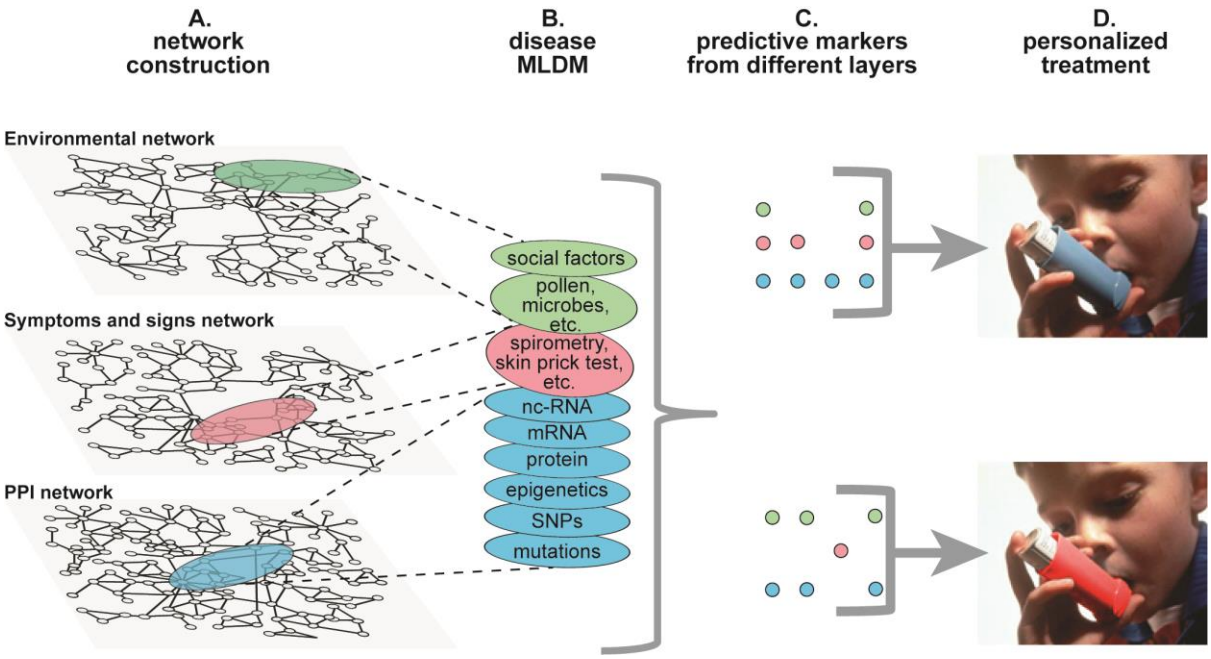


Figure 3. MLDMs for personalised treatment. A) All factors that influence a disease can potentially be described by networks. For example, symptoms and signs, which tend to co-occur can be linked and form a module that correspond to a disease (pink oval). That module may be linked to underlying modular protein changes (blue oval). Similarly, the disease module may be linked to co-occurring environmental factors (green oval). **B**) Each of the modules in A can be further divided to represent different sub-layers, from which **C**) predictive markers from the different sub-layers can be identified, and used for **D**) personalised treatment.

Workshops and conferences

The above discussed content has been presented at the following events:

- CASyM - a roadmap for the implementation of systems medicine in Europe. Key note speaker International Conference in Systems Biology, Copenhagen, September 2013
- Should systems medical training be integrated for clinical and basic researchers? Workshop at International Conference in Systems Biology, Copenhagen, September 2013
- Systems medicine will contribute to 4P Medicine. Lecture at a workshop entitled “Differential Network Medicine”. This was organized for PhD students and post docs organized by a consortium of Belgian universities., Antwerp, September 2013
- Systems medicine will contribute to individualised Medicine. Lecture: Personalized medicine symposium, Karolinska Institute, May 2014
- Key research priorities for systems medicine. Lecture at seminar arranged by the European Commission, Brussels July 2014
- Translational strategy for allergic asthma. Lecture at European Respiratory Society conference in Munich, September 2014
- Omics to identify diagnostic markers in allergic rhinitis. Lecture at European Respiratory Society post doc educational seminar in Munich, September 2014
- Systems medicine to personalise medication. Lecture at the first Systems Biology and Systems Medicine School, Como, September 2014
- Systems medicine – a translational discipline to individualise medicine. Lecture at seminar at the dept of immunology, Gothenburg University, September 2014
- Bioinformatic tools for systems medicine in allergy. Lecture at ERS seminar in Dublin “Human Translational Medicine: A key bridge for development of new drugs for severe asthma, COPD and ILD”
- CASyM Advanced Summer School in System Medicine: Implementation of Systems Medicine across Europe, Sweden 2015, June 2015

References

1. *What happened to personalized medicine?* Nat Biotech, 2012. **30**(1): p. 1-1.
2. Zhang, H., et al., *Targeted omics and systems medicine: personalising care.* Lancet Respir Med, 2014. **2**(10): p. 785-7.
3. Mika Gustafsson, C.E.N., Huan Zhang, Albert-László Barabási, Sergio Baranzini, Sören Brunak, Kian Fan Chung, Howard J Federoff, Anne-Claude Gavin, Richard R Meehan, Paola Picotti, Miguel Àngel Pujana, Nikolaus Rajewsky, Kenneth GC Smith, Peter J Sterk, Pablo Villoslada and Mikael Benson, *Modules, networks and systems medicine for understanding disease and aiding diagnosis.* Genome Med, 2014. **6**(82).
4. Benson, M., *Clinical implications of omics and systems medicine: focus on predictive and individualized treatment.* J Intern Med, 2016. **279**(3): p. 229-40.
5. Jenssen, T.K., et al., *A literature network of human genes for high-throughput analysis of gene expression.* Nat Genet, 2001. **28**(1): p. 21-8.
6. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.
7. Pujana, M.A., et al., *Network modeling links breast cancer susceptibility and centrosome dysfunction.* Nat Genet, 2007. **39**(11): p. 1338-49.
8. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature, 2014. **506**(7488): p. 376-81.
9. Bruhn, S., et al., *A generally applicable translational strategy identifies S100A4 as a candidate gene in allergy.* Sci Transl Med, 2014. **6**(218): p. 218ra4.

ACKNOWLEDGEMENTS

This report is part of CASyM work package 3 – “The technological and methodological basis of systems medicine”.

CASyM is funded by the European Union, Seventh Framework Programme under the Health Cooperation Theme and Grant Agreement # 305033.

STEERING COMMITTEE

The following officials, as part of the Scientific Steering Committee, are involved in the scientific coordination of CASyM:

Charles Auffray - European Institute for Systems Biology & Medicine - EISBM, France
Mikael Benson (Deputy Chair) - Linköping University Hospital, Sweden
Rob Diemel - The Netherlands Organization for Health Research and Development, The Netherlands
David Harrison (Chair) - University of St. Andrews, United Kingdom
Walter Kolch - University College Dublin, Ireland
Frank Laplace - Federal Ministry of Education and Research, Germany
Francis Lévi - Institut National de la Sante et de la Recherche Medicale, France
Damjana Rozman (Deputy Chair) - University of Ljubljana, Faculty of Medicine, Slovenia
Johannes Schuchhardt - MicroDiscovery GmbH, Germany
Olaf Wolkenhauer - Dept. of Systems Biology & Bioinformatics University of Rostock, Germany

ADMINISTRATIVE OFFICE (COORDINATION)

Marc Kirschner - Project Management Jülich, Forschungszentrum Jülich GmbH, Germany


CASYM
EUROPE


SEVENTH FRAMEWORK
PROGRAMME