# Network biology for gene prioritization

Computational biology for human genetics

**Daniela Nitsch**

Jury:
Prof. dr. P. Van Houtte, chairman
Prof. dr. ir. Y. Moreau, promoter
Prof. dr. L. Wehenkel, co-promoter (ULg)
Prof. dr. K. Devriendt
Prof. dr. P. De Causmaecker
Prof. dr. ir. B. De Moor
Prof. dr. phil. S. Brunak (DTU, Copenhagen)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

September 2011

© Katholieke Universiteit Leuven – Faculty of Engineering
Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

Legal depot number D/2011/7515/103
ISBN number 978-94-6018-403-1

# Preface

This thesis contains the research work that I have carried out during my PhD in the Bioinformatics research group of the SISTA research division in the Department of Electrical Engineering (ESAT) at the Katholieke Universiteit Leuven. My research was full of gathering new experiences in research and life, continuous learning and meeting great colleagues and friends. This will always be a grateful memory for my future life.

The most important person behind my thesis and research was my supervisor Prof. Yves Moreau, who I yield my deepest thanks for giving me the opportunity of doing my PhD in his lab on such an exciting topic. I would like to thank him for his continuous support in terms of research questions, funding, travel allowances, and getting in contact with Prof. Søren Brunak who hosted me for a three month period to continue my research in his lab in Denmark.

I also want to acknowledge all the help I received from the administrative staff at SISTA, in particular Ida Tassens, Ilse Pardon, John Vos, Mimi Deprez, and Elsy Vermoesen who were always doing their best for helping me in terms of administrative tasks, traveling or conference attendance.

Beside my supervisor, I want to thank my supervisory committee (Prof. Louis Wehenkel, Prof. Koen Devriendt, and Prof. Bart de Moor) for their participation in my research and for providing fruitful discussions about my research during informal or official meetings. I want to acknowledge Prof. Koen Devriendt in particular for his continuous feedback throughout my PhD in terms of addressing biological questions through computational tools or analyzing results for biological relevance.

For a very helpful collaboration, I want to acknowledge Prof. Marco Saerens from UCL and his team for their support in any machine learning question.
Further, I want to thank Prof. Søren Brunak for hosting me for a three month period at his research group in the Center for Biological Sequence Analysis (CBS) of the Department of Systems Biology at the Technical University of Copenhagen (DTU). There, I have set up a collaborative project together with Tune H. Pers.

During my stay at CBS I could participate in regular activities at the center including group meetings, journal club and close collaborations with other members of the Integrative Systems Biology (ISB) Group. Beside that, I participated at several social events of the Center as well as the yearly CBS Retreat which consists of presentations from each research group of the Center, followed by a social dinner in the evening. I could extend my social network in my research area and my expertise thanks to the expert knowledge available at this computational biology center.

Next, I want to acknowledge my colleagues at BIOI (Olivier Gevaert, Georgios Pavlopoulos, Tunde Adefioye, Jiqiu Cheng, Yousef El Aalamat, Ernesto Iacucci, Arnaud Installe, Peter Konings, Fabian Ojeda, Dusan Popovic, Alejandro Sifrim, Minta Thomas, Léon-Charles Tranchevent, Nico Verbeeck, Lieven Thorrez, Raf Van de Plas, Sonia Leach, Shi Yu) for having fruitful discussions about research and for having pleasant coffee breaks and our weekly lunch together. I further want to remember Anneleen Daemen, Thomas Dhollander, Karen Lemmens, Tim Van den Bulcke, Wouter Van Delm, Roland Bariot, and Sylvain Brohée. In particular, I want to thank Léon-Charles Tranchevent for his continuous support in any research questions, and Tunde Adefioye for his helpful corrections of my thesis. I further want to stress the great help of Lieven Thorrez who was always available when a biological questions arose. I also want to deeply thank Sonia Leach for her kind support during the first year of my PhD in which she introduced the bioinformatics and network biology field to me. It was a pleasure to work on collaborative projects with Léon-Charles Tranchevent and Francisco Bonachela Capdevila. I also want to acknowledge other collaborators: Lieven Thorrez, Amelie Fassbender, Paul Brady, Bernard Thienpont, Prof. Koen Devriendt, and Prof. Hilde Van Esch.

I further want to mention the Bioptrain group (Daniel Soria, Pawel Widera, Enrico Glaab, Andrea Sackmann, Marc Vincent, Linda Fiaschi, Aleksandra Swiercz, and Prof. Jon Garibaldi) for the fruitful exchange of research experience during our pleasant Bioptrain meetings.

Last but not least, I want to thank my beloved parents and brothers as well as my parents-in-law for their continuous support throughout my PhD. In particular, I want to thank my beloved husband Daniel for his patience, support and love throughout the years.

Daniela Nitsch
*Leuven, September 2011*

# Abstract

Discovering novel disease genes is challenging for diseases for which no prior knowledge - such as known disease genes or disease-related pathways - is available. Performing genetic studies frequently result in large lists of candidate genes of which only few can be followed up on for further investigation. In the past couple of years, several gene prioritization methods have been proposed, such as Endeavour, SUSPECT, GeneWanderer, etc. These methods use a guilt-by-association concept (candidate genes that are similar to the already confirmed disease genes are considered promising), and are therefore not applicable when little is known about the phenotype or when no confirmed disease genes are available beforehand.

We have proposed a method that overcomes this limitation by replacing the need for prior knowledge about the biological process with experimental data on differential gene expression between affected and healthy individuals. At the core of the method are a protein interaction network and disease-specific expression data. Our approach propagates the expression data over the network using an extended random walk approach based on kernel methods, as the inclusion of indirect associations compensates for network sparsity.

Candidate genes are ranked based on the differential expression of their network neighborhood. Our method relies on the assumption that strong candidate genes tend to be surrounded by many differentially expressed neighboring genes in a protein interaction network. This allows the detection of a strong signal for a candidate even if its own differential expression value is too small to be detected by a standard analysis, as long as its interacting partners are highly differentially expressed. To assess the performance of this method, we have set up a benchmark. Results showed that it clearly outperforms other gene prioritization approaches with an average ranking position of 8 out of 100 genes, and an AUC value of 92.3% which could lead to promising disease gene discoveries in the future.

To make this method freely available to the community, we have developed a web server, called PINTA, that supports distinct organisms and microarray platforms for candidate gene prioritization.

# Nomenclature

| | |
|---|---|
| BioGRID | Database of protein and genetic interactions |
| Co-IP | Protein complex immunoprecipitation |
| DIP | Database of interacting proteins |
| HPRD | Human protein reference database |
| IntAct | Molecular interaction database |
| iPS cell | induced pluripotent stem cell |
| GEO | Gene expression omnibus |
| GP | Gene prioritization |
| GWAS | Genome-wide association study |
| KEGG | Kyoto encyclopedia of genes and genomes |
| MINT | Molecular interaction |
| MIPS | Mammalian protein-protein interaction database |
| ML | Machine learning |
| OMIM | Online mendelian inheritance in man |
| OPHID | Online predicted human interaction database |
| PCOS | Polycystic ovary syndrome |
| PPI | Protein-protein interaction |
| RW | Random walk |
| STRING | Search tool for the retrieval of interacting genes/proteins |
| TAP purification | Tandem affinity purification |
| T2D | Type 2 diabetes |

# Contents

**Contents** vii

**List of Figures** ix

**List of Tables** xi

**1 Introduction** 1

1.1 Systems biology . . . . . 1

1.1.1 Protein interaction network model . . . . . 2

1.1.2 From networks to graphs . . . . . 5

1.1.3 Random walks on graphs . . . . . 6

1.1.4 Kernels on graphs and distance networks . . . . . 7

1.2 Microarray gene expression profiling . . . . . 11

1.3 Gene prioritization . . . . . 13

1.4 Challenges and objective . . . . . 13

1.5 Structure of the thesis and personal contribution . . . . . 14

**2 *Study 1*: Gene prioritization - a review** 17

**3 *Study 2*: Gene prioritization - a critical assessment** 29

**4 *Study 3*: Gene prioritization by network analysis of expression data** 45

**5 *Study 4*: Gene prioritization by network analysis using machine learning approaches 75**

**6 *Study 5*: PINTA - a web server for network-based gene prioritization 97**

**7 Discussion 103**

7.1 Gene prioritization tools and their critical assessment . . . . . 104

7.2 The purpose of expression based *a priori* gene prioritization . . . . 105

7.3 Case studies & benchmark study . . . . . 106

7.4 The quality and coverage of the underlying network . . . . . 109

7.5 The quality and the relevance of the expression data . . . . . 110

7.6 Monogenic versus polygenic and complex diseases . . . . . 111

7.7 Measuring differential expressed neighborhoods of disease genes . . 111

7.8 PINTA - a web server for gene prioritization . . . . . 112

**8 Perspective 114**

8.1 Using tissue-specific expression data for quality improvement . . . 115

8.1.1 A global gene expression map for tissue-specificity . . . . . 115

8.1.2 Induced pluripotent stem cells for tissue-specificity . . . . . 116

8.2 Data integration for reducing false positives and false negatives . . 118

8.3 Network analysis of time series gene expression data . . . . . 119

8.4 Disease gene discovery for personalized medicine . . . . . 120

8.5 Tissue-specificity of protein-complexes based on expression . . . . 122

8.6 A pipeline of gene prioritization tools for identifying disease genes 124

8.7 Conclusion . . . . . 125

**Bibliography 127**

**List of publications 137**

**Curriculum vitae 141**

# List of Figures

1.1 Exemplary network derived from STRING (Jensen *et al.*) . . . . . 5
1.2 Exemplary network graph . . . . . 6
1.3 Data representation in a feature space . . . . . 7
1.4 Exemplary distance network . . . . . 9
1.5 Structure of the thesis . . . . . 14

7.1 FST subnetwork after gene prioritization (PCOS) . . . . . 108
7.2 CCNO subnetwork after gene prioritization (PCOS) . . . . . 109

# List of Tables

1.1 Protein interactions in the IntAct database . . . . . 3

1.2 Protein-protein interaction databases . . . . . 4

1.3 Graph kernels . . . . . 10

7.1 Gene prioritization results of the polycystic ovary syndrome . . . . 107

8.1 The global map of human expression data for normal and disease tissues after Lukk *et al.* . . . . . 117

8.2 The global map of human expression data for normal after Lukk *et al.* 123

# Chapter 1

# Introduction

This chapter introduces basic concepts of computational biology that represent the core of the studies in this thesis. Section 1 gives an overview of biological networks and basic graph theory in systems biology. Section 2 briefly summarizes the DNA microarray technology, and Section 3 introduces the gene prioritization problem. The challenges and objectives of this thesis are presented in Section 4, followed by its structure and the personal contribution of the PhD candidate in Section 5.

## 1.1 Systems biology

Systems biology is an emerging field aiming at the understanding of biological processes at the system level, focusing on understanding the roles of interactions between genes, proteins, biochemical reactions, and other cell components in an organism. Gene regulation is one aspect of systems biology in which interactions between genes are studied, and how they cause function and behavior of the cell in disease and health. Gene interactions can be revealed as interaction or gene regulatory networks [44].

Although networks play a central role in systems biology, modeling and simulation constitute essential parts of the predictive and explanatory approach to it. Healthcare is an important application for systems biology because models of regulatory networks are useful to understand their alterations in the case of disease and to develop methods to cure the disease. In this context, there is an increasing need for the prediction of systems behavior in drug development, drug validation, diagnostics and therapy monitoring. For complex systems, wiring diagrams can help to understand how components collaborate and how they may cause diseases. By creating a model, the effects of possible perturbations can be predicted in

silico. Furthermore, models gained by systems biology approaches can be used for prediction of the behavior of the biological system even under conditions that are not easily accessible with experiments. Systems biology relies on the integration of experimentation, data processing and modeling in an iterative process [18] [44].

In systems biology, different kinds of networks are established, such as gene networks, protein interaction networks, gene regulatory networks, metabolic networks, signaling networks, etc. In this work, protein interaction networks and functional association networks will be studied to assess the relationship between single genes in the genome.

### 1.1.1 Protein interaction network model

Proteomics focuses on the nature and role of interactions between proteins. Protein-protein interactions (PPIs) have been recognized to be a key element in diverse roles in biology because they regulate a broad range of biological processes, such as transcriptional activation and repression [84], metabolic and developmental control [92], or cell-to-cell interactions [52].

Proteomics experts concluded that most proteins interact with multiple partners and different proteins from intricate interaction networks or highly regulated pathways rather than acting as isolated components [62]. In this context, Phizicky and Fields [21] have analyzed the measurable effects of PPIs as altering the kinetic properties of enzymes, acting as a common mechanism to allow for substrate channeling, creating a new binding site, inactivating, or destroying a protein or changing the specificity of a protein for its substrate through interaction with different binding partners. von Mering and colleagues [95] analyzed annotated proteins revealing that proteins involved in the same cellular process often interact with each other. The function of unknown proteins may be postulated on the basis of their interaction with a known protein target of known function. Mapping PPIs has not only provided insight into protein function but also facilitated the modeling of functional pathways to elucidate the molecular mechanisms of cellular processes. Therefore, studying PPIs is fundamental to further understanding the function of proteins within the cell.

There are distinct technologies to retrieve PPIs in high-throughput studies, such as yeast-two-hybrid assays [88], mass spectrometry [3], protein complex purification [13], protein chip technology [28] [29], etc. These technologies result in a huge amount of heterogeneous PPI data which has to be made coherent for further studies. Despite the existence of databases and integrative PPI networks based on these heterogeneous data sources, it is still challenging to handle this data. Therefore, computational analysis of PPI networks have become a necessary supplemental tool for understanding the functions of uncharacterized proteins (*e.g.,* see [2]).

One exemplary database for storing protein interaction data is the IntAct database [7]. Table 1.1 illustrates the contribution of distinct technologies retrieving PPI data in the IntAct database (as of Feb 14, 2011) and demonstrates that the majority of PPI data was detected by yeast-two-hybrid assays (64.6%) and Coimmunoprecipitation (9.9%).

| Method | number of interactions | Percent |
|---|---|---|
| Two-hybrid | 103,478 | 64.6% |
| Co-IP | 15,928 | 9.9% |
| TAP purification | 5,634 | 3.5% |
| Other | 35,214 | 22.0% |
| Total | 160,254 | 100% |

Table 1.1: The contribution of different PPI methods to protein interactions in the IntAct database (as of Feb 14, 2011)

A PPI network can be described as a complex system of proteins linked by interaction. In the computational analysis of PPI networks, nodes represent proteins or genes. Two proteins or genes physically interacting are represented as adjacent nodes connected by an edge. Based on this graphical representation, various computational approaches, such as data mining, machine learning, and statistical approaches, can be designed to reveal the organization of a PPI network at different levels. For example, neighboring proteins in the network are generally considered to share functions (*guilt-by-association* [9] [36]). Based on this idea, the function of a certain protein may be predicted by considering the functions of those proteins with which it interacts and the complexes to which it belongs. Furthermore, densely connected subgraphs in the network are likely to form protein complexes that function as a unit in a certain biological process [93]. An investigation of the topological features of the network (*e.g.,* whether it is scale-free, a small-world network, or governed by the power law) can also enhance our understanding of the biological system [2] [17]. Recent studies, such as [33] or [75] have attempted to understand and characterize the structural behaviors of PPI networks from a topological perspective. Taylor and colleagues [33], for example, found out that the dynamic structure of the human interactome can be used to predict breast cancer patient outcome. Features, such as small-world properties [17], scale-free distributions [5], and hierarchical modularity [19] have been observed in PPI networks. Therefore, topological methods such as modularity analysis [33] or prediction of protein functions in PPI networks [77] can be applied to further understanding the function of proteins and its assignment in the PPI network.

Databases storing information on PPIs are numerous and diverse. Table 1.2 gives an overview of some well known and widely used databases. Databases, such as MINT (Molecular INTeraction) [1], IntAct (Molecular interaction database) [7], BioGRID (Database of Protein and Genetic Interactions) [8], HPRD (Human Protein

| Database | number of proteins/interactions | number of organisms | URL |
|---|---|---|---|
| DIP [48] | 23,201/71,276 | 372 | http://dip.doe-mbi.ucla.edu |
| MINT [1] | 31,859/90,695 | 30 | http://mint.bio.uniroma2.it/mint |
| IntAct [7] | 56,035/160,254 | 275 | http://www.ebi.ac.uk/intact |
| BioGRID* [8] | 32,951/256,892 | 21 | http://thebiogrid.org |
| HPRD† [91] | 30,047/39,194 | 1 | http://www.hprd.org |
| MIPS [73] | 959/1801 | 10 | http://mips.helmholtz-muenchen.de/proj/ppi |
| STRING‡ [53] | 2,590,259/89,236,924 | 630 | http://string.embl.de |
| OPHID§ [46] | n.a./681,404 | 6 | http://ophid.utoronto.ca |

Table 1.2: Comprehensive databases available for searching and downloading data related to protein interactions (as of Feb 14, 2011). *BioGRID v.3.1.73. † HPRD v.9. ‡ STRING v.8.3. § OPHID v.1.8.

Reference Database) [91], and MIPS (Mammalian Protein-Protein Interaction Database) [73] retrieve their interaction data mainly from literature that was manually curated by experts, whereas the DIP (Database of Interacting Proteins) database [48] catalogs experimentally determined interactions between proteins and combines information from a variety of sources which were curated manually by experts and automatically using computational approaches.

Databases, such as STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [94] [53] or OPHID (Online Predicted Human Interaction Database) [46] combine distinct databases to retrieve interactions and to predict novel interactions or associations between proteins. von Mering and colleagues [94] [53] developed STRING, a database of known and predicted protein interactions which includes direct (physical) and indirect (functional) associations. The associations from which the interaction data is quantitatively integrated are mainly derived from genomic context, high-throughput experiments, (conserved) coexpression, and previous knowledge. The current version of STRING [53] imports protein association knowledge not only from databases of physical interactions, but also from databases of curated biological pathway knowledge such as MINT [1], HPRD [91], BIND [27], DIP [48], BioGRID [8], KEGG [55], Reactome [30], IntAct [7], EcoCyc [32], NCI-Nature Pathway Interaction Database and Gene Ontology (GO) protein complexes. Brown and colleagues [27] introduced OPHID as a database of predicted interactions between human proteins and combines the literature-derived human PPI from BIND, HPRD [91] and MINT [1], with predictions made from distinct organism [46]. In contrast to STRING, OPHID is a PPI network, whereas STRING can be interpreted as a *functional association network* due to the nature of its data sources.

### 1.1.2 From networks to graphs

Depending on the nature of interactions, networks can be either directed or undirected. In directed networks the interaction between two nodes has a direction (*e.g.,* in a transcriptional gene regulatory network the information flow from a collection of regulatory proteins to genes they regulate [89] [82]), while in undirected networks the edges do not have an assigned direction (*e.g.,* in a PPI network edges represent mutual binding relationships, *i.e.,* one protein binds to another protein and vice versa).

In this work, we put a focus on PPI networks which can be represented by undirected graphs in machine learning theory. Nodes in the graph represent proteins and edges represent interactions or associations between two proteins that can be assigned by weights resembling the strength of interaction.

Figure 1.1 shows an exemplary network derived from STRING (v 8.3) [53] in which the edges between the proteins resemble functional associations (STRING can be interpreted as a functional association network as described before).

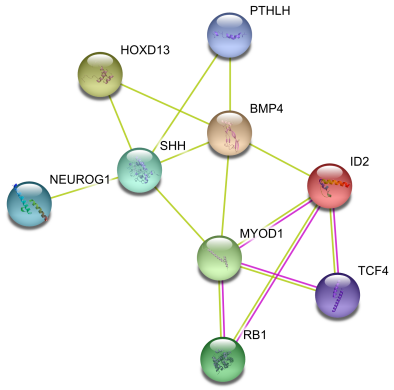



Figure 1.1: Illustration of network derived from STRING (v.8.3) [53] (as of Feb 16 2011)

Consider a given weighted and undirected graph $G = (V, E)$ with symmetric weights $w_{ij} > 0$ between a couple of nodes $v_i$ and $v_j$ that are all linked. The weight $w_{ij}$ increases with the importance of the relation between nodes $v_i$ and $v_j$: the larger its value, the easier the communication through the edge [22]. Figure 1.2 shows the corresponding undirected and weighted graph of the network shown in Figure

1.1, whereas the weights correspond to the evidence scores suggesting a functional link derived from STRING (v.8.3).

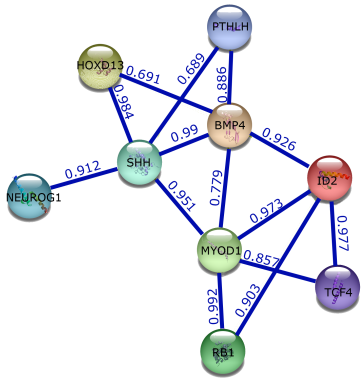



Figure 1.2: The weighted and symmetric graph corresponding to the network derived from STRING and shown in Figure 1.1.

The Adjacency matrix $A$ is defined as $a_{ij} = w_{ij}$ if the nodes $v_i$ and $v_j$ are directly connected and $a_{ij} = 0$ otherwise. The Laplacian matrix $L$ of $G$ is defined as $L = D - A$ , with the diagonal degree matrix $D = diag(a_{i\cdot})$ with entries $d_{ii} = a_{i\cdot} = \sum_{j=1}^{n} a_{ij}$. The Laplacian matrix $L$ of graph $G$ with eigenvalues $\lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{n-1}$ is always symmetric and positive semidefinite (i.e., $\forall i, \lambda_i \geq 0$). If $L$ is fully connected, it has rank $n - 1$ [24] [78].

### 1.1.3 Random walks on graphs

A random walk (RW) is a mathematical formalization of a trajectory that consists of taking successive random steps. Let $G = (V, E)$ be a connected graph with $v_0, ..., v_n$ nodes. Performing a random walk on $G$ means starting at a node $v_0$ and arriving at node $v_t$ after $t$ steps. The sequence of nodes visited by a random walker is called a *Markov chain* or *random walk* [22]. Let the random walker be at node $v_i$ at time $t$ with the random variable $s(t)$ containing the current node of the Markov chain at time $t$: $s(t) = v_i$. Then, a random walk can be specified with single-step transition probabilities of jumping from any node $v_i$ to an adjacent node $v_j$:

$$v_j = s(t+1) : P(s(t+1) = v_j | s(t) = v_i) = \frac{a_{ij}}{ai\cdot} = p_{ij}. \tag{1.1}$$

The transition probabilities $p_{ij}$ only depend on the current node $v_i$ and not on the past ones [22] [47]. Lovasz [47] assumes a random walk to be a Markov chain or

Markov process that is time-reversible whose basis properties are determined by the spectrum of the graph.

A Markov chain is symmetric (*i.e.,* the probability of moving from node $i$ to node $j$ is equal to the probability of moving from node $j$ to $i$) if $G$ is regular [1], and the random walk is *stationary* because the distribution on $V$ is uniform (*i.e.*, $P(s(t)) = P(s(0))$ for all $t \geq 0$) [47]. Reversely, if $G$ is not regular, time-reversibility is held and a random walk considered backwards is also a random walk [47]. In practice, if a PPI network (such as OPHID [46]) is represented by an undirected graph on which a random walk is executed, the walk is a time-reversible Markov chain. Since PPI networks hold the small-world property [17], the underlying graph is not regular and the Markov chain is not symmetric.

### 1.1.4 Kernels on graphs and distance networks

Kernel methods are widely used algorithms for pattern analysis, such as clustering, classification, correlation, regression or discriminant analysis. Working in a kernel-defined feature space means not to explicitly represent points, but instead accessing to the evaluation of inner products between points. Let $\Phi$ be an embedding map from $X$ to an (inner product) feature or Hilbert space $F$ [37]:

$$\Phi : x \in \mathbb{R}^n \longmapsto \Phi(x) \in F \subseteq \mathbb{R}^n. \tag{1.2}$$

A kernel is a function $\kappa$ for any two points $x, z \in X$ that corresponds to an inner product or dot product in $F$:

$$\kappa(x, z) = \langle \Phi(x), \Phi(z) \rangle . \tag{1.3}$$

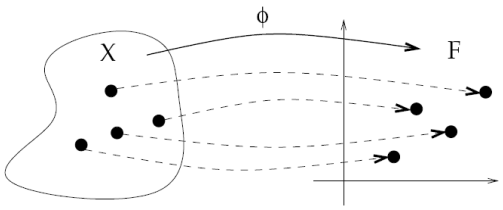



Figure 1.3: For any kernel $k$ on a space $X$, there exists a feature space $F$ and a mapping $\phi : X \to F$ [37]

The corresponding kernel matrix $K$ is defined as

$$K_{ij} = \kappa(x_i, z_j) \text{ for } i, j = 1, ..., l. \tag{1.4}$$

[1] "*A Markov chain is a random walk on a weighted directed graph. Similarly, time-reversible Markov chains are random walks on undirected graphs, and symmetric Markov chains are random walks on regular symmetric graphs.*" [47]

A kernel function $\kappa : F \times F \rightarrow \mathbb{R}$ returns a real number $\kappa(x, z)$ for two points $x$ and $z$ in $F$. The inner product of $x$ and $z$ is a classical similarity measure assuming that they are expressed in a vector space. For kernel functions $\kappa$ the computation of $\kappa(x, z)$ is equal to the computation of the inner product of $x$ and $z$ in another space [37].

Kernel-based algorithms allow for computing inner products (*i.e.,* similarities) in a high-dimensional feature space and similarities between structured objects that cannot be naturally represented by a simple set of features. A kernel is expected to capture an appropriate measure of similarity for a particular task and to require significantly less computation than would be needed via an explicit evaluation of the corresponding mapping from input space into feature space [23].

A valid kernel function is symmetric and positive semidefinite and holds the following definitions [23]: (1) $x'Kx \geq 0$ for all $x \in \mathbb{R}^n$, (2) all Eigenvalues of $K$ are positive, (3) $K$ is the inner product of linearly independent vector $x_1, ..., x_n \in \mathbb{R}^r$ for some $r$, and (4) $K$ is a diagonal matrix $\Lambda$ in another coordinate system $K = U\Lambda U'$.

Furthermore, Fouss and colleagues [23] showed that the kernel-based similarity measures is Euclidean (*i.e.,* the nodes of the graph can be embedded in an Euclidean space preserving distances between nodes). To prove that they took all direct and indirect paths between graph nodes into account for kernel-based similarity measures. For increasing number of paths between two nodes and decreasing of the length of any path, the similarity measure increases. Thus, the more short paths connected two nodes, the more similar those nodes are.

Consider, for example, the Gaussian radial basis function (RBF) kernel:

$$k_G(x, z) = \exp\left(-\frac{d(x, z)^2}{2\sigma^2}\right), \tag{1.5}$$

where $\sigma$ is a parameter and $d$ is the Euclidean distance. This is a valid kernel $k_G(x, z) = \langle\phi(x), \phi(z)\rangle$ which is a decreasing function of the Euclidean distance between points, and therefore has a relevant interpretation as a measure of similarity: the larger the kernel $k_G(x, z)$, the closer the points $x$ and $z$ in $X$ [37]. This so called *kernel trick* or *kernalization* is a principle based on the property of kernels discussed before: Any algorithm for vectorial data that can be expressed only in terms of dot products between vectors can be performed implicitly in the feature space associated with any kernel, by replacing each dot product by a kernel evaluation [37].

Illustrating this principle on computing distances between two points $x$ and $z$ they are mapped to two vectors $\phi(x)$ and $\phi(z)$ in $F$, so a distance $d(x, z)$ between the objects are defined as the Hilbert distance between their images,

$$d(x, z) := ||\phi(x) - \phi(z)||. \tag{1.6}$$

The distance can be expressed in terms of dot products in $F$:

$$||\phi(x) - \phi(z)||^2 = \langle \phi(x), \phi(x) \rangle + \langle \phi(z), \phi(z) \rangle - 2 \langle \phi(x), \phi(z) \rangle . \tag{1.7}$$

Applying the kernel trick in Equation 1.7 and plugging results into Equation 1.6 shows that the distance can be computed only in terms of the kernel [37],

$$d(x, z) = \sqrt{\kappa(x, x) + \kappa(z, z) - 2\kappa(x, z)}. \tag{1.8}$$

In *Study 3* and *Study 4*, we have called the resulting kernel matrix that we have computed from a PPI or functional association network respectively, *distance network*, highlighting its natural distance measure [23] as described before. The purpose of computing a distance network was to determine distances based on a random walk between all pairs of genes even if they were not directly interacting or associated with each other but rather indirectly via other genes.

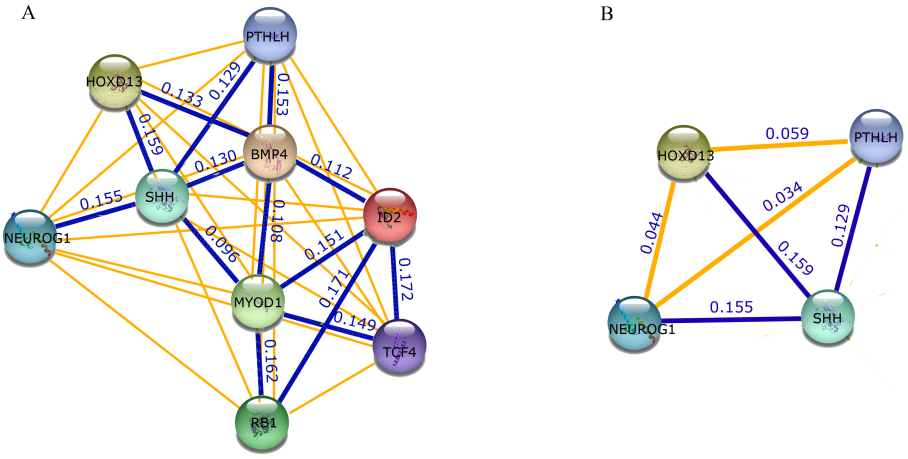



Figure 1.4: Distance network after computing the Laplacian Exponential Diffusion kernel using the network shown in Figure 1.1 and Figure 1.2. A presents the complete distance network with distances for all directly connected node pairs (as shown in Figure 1.2), whereas B presents a small subnetwork of A with all computed distances between four nodes. Blue edges represent directly connected nodes as derived from Figure 1.2, and orange edges represent indirect connections between nodes.

Figure 1.4 shows a distance network after computing the Laplacian Exponential Diffusion kernel [78]. Figure 1.4-A presents the complete distance network with nodes connected by blue edges representing direct connections as derived from the graph in Figure 1.2, and nodes connected by orange edges showing that those

nodes do not have any direct but rather indirect connection with each other in the graph in Figure 1.2. However, distances can be computed between all pairs of nodes, independent of whether the interaction or association is direct or indirect. Figure 1.4-B presents a small subnetwork of the distance network in Figure 1.4-A showing all distances between four nodes.

In Equation 1.5 a valid kernel function was already introduced. However, in the literature, several distinct kernel functions can be found (for an overview see Table 1.3).

| Graph kernel | Definition | Ref. |
|---|---|---|
| Exponential Diffusion kernel | $K_{ED} = exp(\alpha A)$ | [67] |
| Laplacian Exponential Diffusion kernel | $K_{LED} = exp(\alpha L)$ | [78] |
| von Neumann Diffusion kernel | $K_{VND} = (I - \alpha A)^{-1}$ | [67] |
| Regularized Laplacian kernel | $K_{RL} = (I + \alpha L)^{-1}$ | [87] |
| Commute-Time kernel | $K_{CT} = L^{+}$ | [58] |
| Regularized Commute-Time kernel | $K_{RCT} = (D - \alpha A)^{-1}$ | [23] |
| Random-Walk-With-Restart similarity matrix | $K_{RWR} = (D - \alpha A)^{-1} D$ | [23] |

Table 1.3: Graph kernels

Kandola and colleagues [67] introduced the Exponential Diffusion kernel with $\alpha$ as the diffusion parameter that determines the degree of diffusion.The kernel integrates a contribution from all paths connecting node $v_i$ and node $v_j$, discounting paths according to their number of steps. It favors shorter paths between two nodes by giving them a heavier weight. The discounting factor is $\frac{\alpha^k}{k!}$. A diffusion process on a graph can be seen as a model of semantic relations existing between indirectly connected terms. Although the number of possible paths between two nodes can increase exponentially, it was shown that it is possible to compute the similarity between two nodes efficiently without examining all possible paths.

Kondor and Lafferty [78] introduced the Laplacian Exponential Diffusion kernel that is similar to the Exponential Diffusion kernel, but based on the Laplacian matrix. The kernel matrix $K$ can be seen as a random walk, starting from a node and transitioning to a neighboring node with the probability $\alpha$. Fixing the transition probability to every neighboring node as $\alpha$, the probability of staying at the current node $i$ is 1-$d_i\alpha$, where $d_i$ is the degree of node $i$. The column vector $j$ of $K$ then represents the steady-state probability vector of the random walk when starting at node $j$, and the value $K_{ij}$ represents the probability that a random walk starting from $i$ will be at $j$ after infinite time steps. If the diffusion parameter is small, $K$ can be seen as a lazy random walk because the probability that the random walker stays at the current node is high. As $\alpha$ increases, the kernel values diffuse more completely through the graph, and when $\alpha$ is sufficiently large, the values among distant nodes capture the long-range relationships between nodes.

Kandola and Lafferty [67] introduced the von Neumann Diffusion kernel for computing document similarity from terms occurring in documents. It is defined in terms of the document-by-term matrix $X$ whose $(i, j)$-element is the frequency of the $j$-th term occurring in document $i$.

Ito and colleagues [87] introduced the regularized Laplacian kernel that differs from the von Neumann diffusion kernel only by substituting the negated Laplacian matrix -L for the adjacency matrix $A$. Following [87] this kernel can be interpreted as path counting, which takes place in the graph induced by the Laplacian matrix instead of the co-citation graph itself. Unlike other kernels (*e.g.,* von Neumann kernel), this kernel assigns negative weights to self-loops in order to remain a relatedness measure for all values of $\alpha$.

Saerens and colleagues [24] introduced the average commute time as the Commute-Time kernel as the Moore-Penrose pseudoinverse of the Laplacian matrix of the graph. The elements of $L^+$ are inner product of the node vectors in the Euclidean space where these node vectors are exactly separated by commute time distances. The Commute-Time kernel takes its name from the average commute time $n(v_i, v_j)$, which is the average number of steps that a random walker, starting in node $v_i \neq v_j$, takes before entering node $v_j$ for the first time and then going back to $v_i$ [58].

The Regularized Commute-Time kernel was introduced by Fouss and colleagues [23] with the parameter $\alpha$ being between 0 and 1. Since this kernel is the matrix inverse of the sum $((1 - \alpha)D + \alpha L)$ of a positive definite matrix and a positive semidefinite matrix, it is positive definite and a valid kernel.

The Random-Walk-With-Restart similarity matrix was introduced by Fouss and colleagues [23]. Like the diffusion kernel, the model considers a random walker jumping from some node $v_i$ and to some neighbor node $v_j$ with a probability proportional to the edge weight $p_{ij}$ as defined in Equation 1.1. In addition, at each step of the random walk, the random walker has some probability $(1 - \alpha)$ to return to node $v_i$ instead of continuing to neighbor nodes [23].

## 1.2 Microarray gene expression profiling

DNA microarrays belong to a well established technology in life science. They contain a large number of DNA samples for simultaneously detecting and quantifying of thousands of different cDNA transcripts in a single experiment. Different platforms for distinct applications have been developed, such as cDNA microarrays for gene expression profiling [59], oligonucleotide arrays for high-throughput SNP genotyping [79], comparative genomic hybridization (array-CGH) for DNA copy number detection [14] [76], chromatin immunoprecipitation on chip (ChIP-on-chip) for investigating interactions between proteins and DNA in vivo [45], etc.

In this work, a main focus will be set on cDNA microarrays for obtaining expression levels of genes or DNA fragments.

In general, gene expression profiling is performed to assess the difference in gene expression level between two conditions (*e.g.,* control vs. case) and is called *case-control study*. As an example, this study could analyze a healthy individual in comparison with a disease individual (*e.g.,* cancer patient) or an individual under a certain condition (*e.g.,* drug treated patient). It allows for detection of genes that are *differentially expressed* in case vs. control and helps researchers to follow up the question which genes could be involved in the disease under study due to their upregulated expression in a healthy person and downregulated expression in a disease person or vice versa.

There are distinct ways to determine the differential expression $DE$ for a gene $g$ using its expression values $e_g$ over all microarray experiments under study, such as the ratio between both conditions:

$$DE_R(g) = \frac{case(e_g)}{control(e_g)}. \tag{1.9}$$

The log2-ratio (or fold change) considers genes that differ by more than an arbitrary cut-off value to be differentially expressed (*e.g.,* for a two-fold difference as the cut-off value genes are taken to be differentially expressed if the expression under one condition is over two-fold greater or less than that under the other condition):

$$DE_{LR}(g) = log_2(\frac{case(e_g)}{control(e_g)}). \tag{1.10}$$

The t-test statistic illustrates the difference between the group means (signal) and variability of groups (noise) and determines which genes are significantly differentially expressed:

$$DE_t = \frac{\mu_{case(e_g)} - \mu_{control(e_g)}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \tag{1.11}$$

A modified version of the t-test is the CyberT [72][97], that uses a Bayesian estimate of the variance among gene measurements within an experiment.

To determine the significance of a gene's differential expression value, *p-values* can be computed based on the t-test statistic, the modified t-test statistic or by permutation analysis (permuting the expression values and randomly reassigning the expression values to the genes). This allows to measure the evidence against the null hypothesis and determines the probability of the occurrence of a test statistic equal to, or more extreme than, the observed value under the assumption that the null hypothesis is true [83].

## 1.3 Gene prioritization

The identification of novel disease genes or genetic variants underlying genetic disorders for effective diagnostic testing and elaborating novel treatments for those diseases still remains a major challenge in human genetics. In the past, high-throughput technologies, such as association studies and linkage analyses, could reveal discoveries in this field, such as identifying chromosomal regions involved in the phenotype of interest [76] [66]. However, these technologies mainly return large lists of candidate genes of which only one or a few are really associated with the disease or phenotype under study [51]. Identifying, among such a list, the most promising candidate genes is a challenging task for biologists because they have to manually go through the list and check whether a gene can be assumed to be promising or not. This task has been introduced by the bioinformatics community as *gene prioritization*. It has been developed into a key in genetics because it is generally too expensive and time-consuming to experimentally validate all candidate genes and to handle the large amount of genomic data that has been made publically available until nowadays. Since then, many distinct computational approaches have been developed to prioritize genes for genetic diseases (*e.g.,* [20] [39] [81] [34] [15]). Most of them follow the guilt-by-association concept, which means that candidate genes that are similar to the already confirmed disease genes are considered promising [9]. The approaches mainly differ in the data sources they use, such as literature, functional annotations, interactions, expression data, sequence information, etc. To find a way through this maze of gene prioritization approaches and their applications, several reviews and studies have been performed to assess their performance in real applications [57] [69] [63].

## 1.4 Challenges and objective

Most of the available gene prioritization methods, such as SUSPECT [20], Endeavour [81] [50], or ToppGene [34], require prior knowledge about the disease to identify putative disease genes, for example in the form of a set of genes, pathways, or gene ontology categories known to be implicated in the disease. However, for unknown diseases or novel phenotypes these methods will be difficult to apply because of the lack of information. Leading to a tendency of biologists and clinicians using a standard genetic procedure based on experimental data to analyze candidate genes by assessing their expression levels in patient-derived material against healthy controls. Candidates for which a significant difference is observed between these two groups are considered promising. However, for many genetic diseases (such as diseases arising from point mutations in coding regions), there is no guarantee that the expression level of the disease gene itself is affected.

In this work we want to overcome these limitations by replacing prior knowledge with experimental data on differential gene expression between affected (patients) and healthy (controls) individuals, and by considering the differential expression data at the level of a gene network rather than in isolation. We assess a candidate gene by considering the level of differential expression in its neighborhood under the assumption that strong candidates will tend to be surrounded by differentially expressed neighbors.

## 1.5 Structure of the thesis and personal contribution

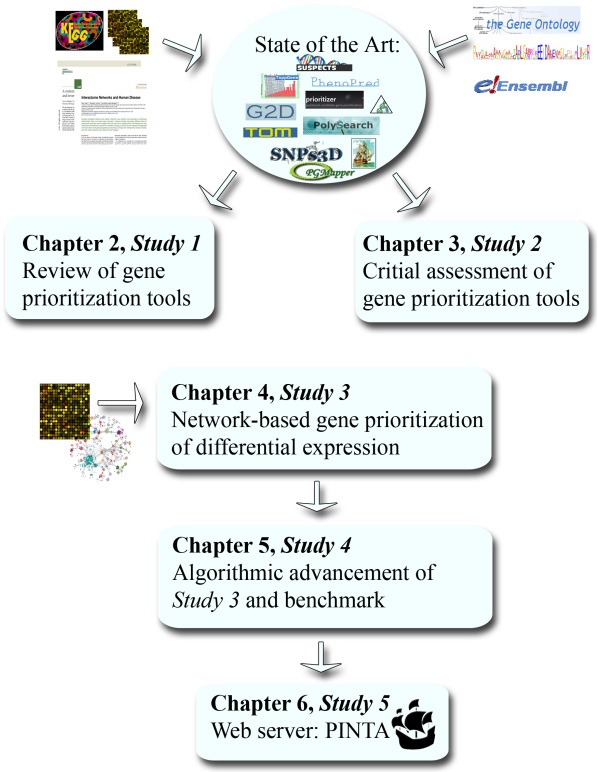



Figure 1.5: Structure of the thesis.

Chapter 2 and 3 give an overview of the state of the art in the gene prioritization field and computational approaches that have been made available by the bioinformatics community to address the gene prioritization challenge. Here, we report two studies: the first one (*Study 1*, Chapter 2) provides a review of gene prioritization tools that

are freely accessible via web interfaces, aiming at providing a guide for biologists through the maze of gene prioritization methods and showing how they differ from each other in terms of input and output data. Secondly, *Study 2* (Chapter 3) assesses the performance of these tools by evaluation of a benchmark data set based on distinct diseases.

Chapter 4-6 present three studies in which a novel approach has been developed for *a priori* candidate gene prioritization when nothing is known about the disease or phenotype under study by replacing prior knowledge with experimental data on differential gene expression between patients and controls. In the first study (*Study 3*, Chapter 4) we have introduced our concept and presented a first implementation evaluated on four distinct diseases. The follow up study (*Study 4*, Chapter 5) presents an algorithmic advancement and optimization of the previous method evaluated by a benchmark on mouse knockout experiments. Finally, we have implemented a web server in *Study 5* (Chapter 6), namely PINTA, to make our method available for the user.

Chapter 7 describes the strengths and limitations of our method followed by a detailed discussion.

Chapter 8 addresses limitations of our method discussed in Chapters 4-6 and provides potential conceptual solutions that could be applied in the future. Furthermore, two ongoing projects are presented here which will be finished up in the near future.

**Personal contribution**
*Study 1* represents a collaborative effort with equal contribution from the PhD candidate and two co-authors in terms of developing the idea, setting up the study, reviewing the tools, performing the experiments, analyzing the data and writing the paper. *Study 2* represents a collaborative effort with equal contribution from the PhD candidate and two co-authors in terms of developing the idea, setting up the study, performing the experiments, analyzing the data and writing the paper. The idea of *Study 3* was developed by the PhD candidate and her supervisor. The PhD candidate has designed and implemented the method in Matlab, set up and applied the case studies for evaluating the method, analyzed the results and drafted the paper. In *Study 4* the PhD candidate has designed and implemented the algorithms in Matlab, designed and set up the benchmark data set, preprocessed and normalized the experiments, analyzed the results and drafted the paper. The web server in *Study 5* is based on the Matlab implementation of *Study 4* and was designed and implemented by the PhD candidate. The PhD candidate further wrote the paper.

# Chapter 2

# *Study 1*: Gene prioritization - a review

In the past, many distinct computational approaches have been developed to prioritize genes for genetic diseases. Although several reviews and studies have been performed to assess their performance in real applications [57] [69] [63], we aim at providing a guide for biologists through the maze of gene prioritization methods that represent the state of the art and that are available as web applications. Here, we show how they differ from each other in terms of input required and output provided, as well as the databases and information sources that are integrated. In addition, we have developed a web site, namely *Gene Prioritization Portal*, representing an updatable electronic version of this review.

**Publication**
This work has been published in 2010 as an electronic version and in 2011 as a printed version of the *Briefings in Bioinformatics* journal (ISI IF 2009: 7.3).
*Reference*: Tranchevent L*, Capdevila FB*, Nitsch D*, De Moor B, De Causmaecker P, Moreau Y (2011). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics* 12(1):22-32. Published online March 21, 2010 doi:10.1093/bib/bbq007. PMID:21278374.
*Equally contributed.

**Personal contribution**
This work represents a collaborative effort with equal contribution from the PhD candidate and two co-authors in terms of developing the idea, setting up the study, reviewing the tools, performing the experiments, analyzing the data and writing the paper.

Briefings in Bioinformatics Advance Access published March 21, 2010

doi:10.1093/bib/bbq007

# A guide to web tools to prioritize candidate genes

*Léon-Charles Tranchevent*, Francisco Bonachela Capdevila*, Daniela Nitsch*, Bart De Moor, Patrick De Causmaecker and Yves Moreau*

Submitted: 8th January 2010; Received (in revised form): 8th February 2010

## Abstract

Finding the most promising genes among large lists of candidate genes has been defined as the gene prioritization problem. It is a recurrent problem in genetics in which genetic conditions are reported to be associated with chromosomal regions. In the last decade, several different computational approaches have been developed to tackle this challenging task. In this study, we review 19 computational solutions for human gene prioritization that are freely accessible as web tools and illustrate their differences. We summarize the various biological problems to which they have been successfully applied. Ultimately, we describe several research directions that could increase the quality and applicability of the tools. In addition we developed a website (http://www.esat.kuleuven.be/gpp) containing detailed information about these and other tools, which is regularly updated. This review and the associated website constitute together a guide to help users select a gene prioritization strategy that suits best their needs.

***Keywords:*** *gene prioritization; candidate gene; disease gene;* in silico *prediction; review*

## BACKGROUND

One of the major challenges in human genetics is to find the genetic variants underlying genetic disorders for effective diagnostic testing and for unraveling the molecular basis of these diseases. In the past decades, the use of high-throughput technologies (such as linkage analysis and association studies) has permitted major discoveries in that field [1, 2]. These technologies can usually associate a chromosomal region with a genetic condition. Similarly, one can also use expression arrays to obtain a list of transcripts differentially expressed in a disease sample with respect to a reference sample. A common characteristic of these methods is usually the large size of the chromosomal regions returned, typically several megabases [3]. The working hypothesis is often that only one or a few genes are really of primary interest (i.e. causal). Identifying the most promising candidates among such large lists of genes is a challenging and time consuming task. Typically, a biologist would have to go manually through the list of candidates, check what is currently known about

Corresponding author. Yves Moreau, Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium. Tel: +32 (0)16 32 8645; Fax: +32 (0)16 32 1970; E-mail: yves.moreau@esat.kuleuven.be

*These authors contributed equally to this work.

**Léon-Charles Tranchevent** is a PhD student at the Katholieke Universiteit Leuven. His main research topic is the development of computational solutions for the identification of disease causing genes through the fusion of multiple genomic data.

**Francisco B. Capdevila,** is a PhD student at the Katholieke Universiteit Leuven. His main research interest is the application of machine learning techniques, specially clustering, in gene prioritization.

**Daniela Nitsch** is a PhD student at the Katholieke Universiteit Leuven. Her research focus on the identification of disease causing genes through the exploration of gene and protein network based techniques.

**Bart De Moor** is a full Professor at the Department of Electrical Engineering of the Katholieke Universiteit Leuven. His research interests are in numerical linear algebra and optimization, system theory and system identification, quantum information theory, control theory, data-mining, information retrieval and bioinformatics.

**Patrick De Causmaecker** is an Associate Professor at the Department of Computer Science at the Katholieke Universiteit Leuven, Head of the CODeS Research Group on Combinatorial Optimisation and Decision Support.

**Yves Moreau** is a Professor at the Department of Electrical Engineering and a Principal Investigator of the *SymBioSys* Center for Computational Systems Biology of the Katholieke Universiteit Leuven. His two main research themes are the development of (i) statistical and information processing methods for the clinical diagnosis of constitutional genetic and (ii) data mining strategies for the identification of disease causing genes from multiple omics data.

© The Author 2010. Published by Oxford University Press. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

each gene, and assess whether it is a promising candidate or not. The bioinformatics community has therefore introduced the concept of gene prioritization to take advantage of both the progress made in computational biology and the large amount of genomic data publicly available. It was first introduced in 2002 by Perez-Iratxeta *et al.* [4] who already described the first approach to tackle this problem. Since then, many different strategies have been developed [5–34], among which some have been implemented into web applications and eventually experimentally validated. A similarity between all strategies is their use of the 'guilt-by-association' concept: the most promising candidates will be the ones that are similar to the genes already known to be linked to the biological process of interest [35–37]. For example, when studying type 2 diabetes (T2D), KCNJ5 appears as a good candidate through its potassium channel activity [38], an important pathway for diabetes [39], and because it is known to interact with ADRB2 [40], a key player in diabetes and obesity. This notion of similarity is not restricted to pathway or interaction data but rather can be extended to any kind of genomic data. Recently, initial efforts have been made to experimentally validate these approaches. For instance, in 2006, two independent studies used multiple tools in conjunction to propose new meaningful candidates for T2D and obesity [41, 42]. More recently, Aerts *et al.* [43] have developed a computationally supported genetic screen whose computational part is based on gene prioritization (Figure 1).

With this review, we aim at describing the current options for a biologist who needs to select the most promising genes from large candidate gene lists. We have selected strategies for which a web application was available, and we describe how they differ from each other and, when applicable, how they were successfully applied to real biological questions. In addition, since it is likely that novel methods will be proposed in the near future, we have also developed a website termed 'Gene Prioritization Portal' (available at: http://www.esat.kuleuven.be/gpp/) that represents an updatable electronic review of this field.

## SELECTING THE GENE PRIORITIZATION TOOLS

In this study, we review 19 gene prioritization tools that fulfill the two following criteria. First, the strategy should have been developed for human candidate disease gene prioritization. Notice that predicting the function of a gene or its implication in a genetic condition are two closely related problems. Moreover, several gene function prediction methods have indeed been applied to disease gene prioritization with reasonable performance [5]. However, it has been shown that gene prioritization is more challenging than gene function prediction since diseases often implicate a complex set of cascades covering different molecular pathways and functions [44]. Besides, to our knowledge, none of the existing gene function prediction methods includes disease-specific data. Thus, these methods were excluded from the present study. For gene function prediction methods, readers are referred to the reviews by Troyanskaya *et al.* [45] and Punta *et al.* [46]. Our second criterion is that a functional web application should be available for the proposed strategy. Since the end users of these tools are not expert in computer science, approaches only providing a set of scripts, or some code to download have been discarded. Furthermore, we focus our analysis on the noncommercial solutions and thus require the web tools to be freely accessible for academia. Using these criteria, we were able to retain a total of 19 applications that still differ by (i) the inputs they need from the user, (ii) the computational methods they implement, (iii) the data sources they use and (iv) the output they present to the user. The thorough discussion of these characteristics has allowed us to create a decision tree (Figure 2) that supports users in their decision process.

In the following section, we summarize the gene prioritization tools that we have retained. The corresponding references and the URL of their web applications are presented in Table 1. Several approaches combine different data sources. SUSPECT ranks candidate genes by matching sequence features, gene expression data, Interpro domains, and GO terms [6]. CANDID uses several heterogeneous data sources, some of them chosen to overcome bias [7]. Endeavour is, however, using training genes known to be involved in a biological process of interest and ranks candidate genes by applying several models based on various genomic data sources [8].

Among the tools using different data sources, ToppGene, SNPs3D, GeneDistiller and Posmed include mouse data within their algorithms, but in a different manner. ToppGene combines mouse phenotype data with human gene annotations and literature [9]. SNPs3D identifies genes that are candidates for being involved in a specified disease based on literature [10]. GeneDistiller uses mouse

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

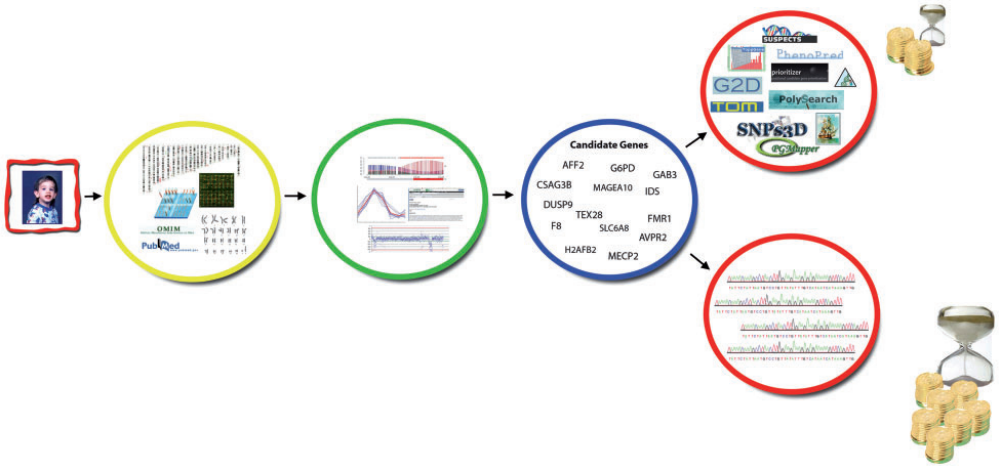



**Figure 1:** A major challenge in human genetics is to unravel the genetic variants and the molecular basis that underlay genetic disorders. In the past decades, geneticists have mainly used high-throughput technologies (such as linkage analysis and association studies). These technologies usually associate a chromosomal region, possibly encompassing dozens of genes, with a genetic condition. Identifying the most promising candidates among such large lists of genes is a challenging and time consuming task. The use of computational solutions, such as the ones reviewed in that paper, could reduce the time and the money spent for such analysis without reducing the effectiveness of the whole approach.

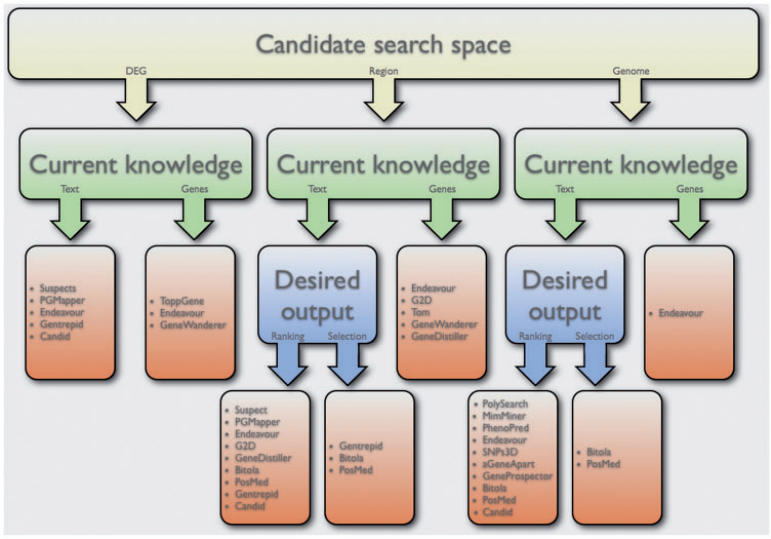



**Figure 2:** Decision tree that categorizes the 19 gene prioritization tools according to the outputs they use and the outputs they produce. This tree is designed to support the end users in their decision so that they can choose the tools that suit best their needs. By starting from the first question on the top and by going down, the user can determine a list of tools that can be used; in addition, the Figure 3 that describes the data sources used by the tool can also be used to support the decision.

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

**Table 1:** Overview of the 19 tools reviewed in the current study with their corresponding publications and website

| Tool | References | Website |
|---|---|---|
| SUSPECT | [6] | http://www.genetics.med.ed.ac.uk/suspects/ |
| ToppGene | [9] | http://toppgene.cchmc.org/ |
| PolySearch | [15] | http://wishart.biology.ualberta.ca/polysearch/index.htm |
| MimMiner | [16] | http://www.cmbi.ru.nl/MimMiner/cgi-bin/main.pl |
| PhenoPred | [23] | http://www.phenopred.org |
| PGMapper | [21] | http://www.genediscovery.org/pgmapper/index.jsp |
| Endeavour | [8, 32] | http://www.esat.kuleuven.be/endeavour |
| G2D | [33, 34] | http://www.ogic.ca/projects/g2d_2/ |
| TOM | [13, 14] | http://www-micrel.deis.unibo.it/~tom/ |
| SNPs3D | [10] | http://www.SNPs3D.org |
| GenTrepid | [20] | http://www.gentrepid.org/ |
| GeneWanderer | [22] | http://compbio.charite.de/genewanderer |
| Bitola | [17] | http://www.mf.uni-lj.si/bitola/ |
| CANDID | [7] | https://dsgweb.wustl.edu/hutz/candid.html |
| PosMed | [12] | http://omicspace.riken.jp |
| GeneDistiller | [11] | http://www.genedistiller.org/ |
| aGeneApart | [18] | http://www.esat.kuleuven.be/ageneapart |
| GeneProspector | [19] | http://www.hugenavigator.net/HuGENavigator/geneProspectorStartPage.do |

phenotype to filter genes [11] and Posmed utilizes among other data sources orthologous connections from mouse to rank candidates [12].

G2D uses three algorithms based on different prioritization strategies to prioritize genes on a chromosomal region according to their possible relation to an inherited disease using a combination of data mining on biomedical databases and gene sequence analysis [4]. TOM efficiently employs functional and mapping data and selects relevant candidate genes from a defined chromosomal region [13, 14].

Tools that are mainly based on literature and text mining are PolySearch, MimMiner, BITOLA, aGeneApart and GenePropector. PolySearch extracts and analyses relationships between diseases, genes, mutations, drugs, pathways, tissues, organs and metabolites in human by using multiple biomedical text databases [15]. MimMiner analyses the human phenome by text mining to rank phenotypes by their similarity to a given disease phenotype [16] and BITOLA mines MEDLINE database to discover new relations between biomedical concepts [17]. aGeneApart creates a set of chromosomal aberration maps that associate genes to biomedical concepts by an extensive text mining of MEDLINE abstracts, using a variety of controlled vocabularies [18]. GeneProspector searches for evidence about human genes in relation to diseases, other phenotypes and risk factors, and selects and prioritizes candidate genes by using a literature database of genetic association studies [19].

Finding associations between genes and phenotypes is the focus of Gentrepid and PGMapper.

Whereas Gentrepid predicts candidate disease genes based on their association to known disease genes of a related phenotype [20], PGMapper matches phenotype to genes from a defined genome region or a group of given genes by combining the mapping information from the Ensembl database and gene function information from the OMIM and PubMed databases [21].

Tools, such as GeneWanderer, Prioritizer, Posmed and PhenoPred, make use of genomewide networks. GeneWanderer is based on protein–protein interaction and uses a global network distance measure to define similarity in protein–protein interaction networks [22]. PhenoPred uses a supervised algorithm for detecting gene–disease associations based on the human protein–protein interaction network, known gene–disease associations, protein sequence and protein functional information at the molecular level [23]. Instead of using a human protein–protein interaction network, Posmed is based on an artificial neural network-like inferential process in which each mined document becomes a neuron (documentron) in the first layer of the network and candidate genes populate the rest of layers [12].

Although we have limited our analysis to the tools freely accessible via a web interface, we are aware of other gene prioritization methods that were excluded of the present analysis but that still represent important contributions to the field. First,

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

**Box 1:** Glossary

**Gene prioritization**
The gene prioritization problem has been defined as the identification of the most promising candidate genes from a large list of candidates with respect to a biological process of interest.

**Data sources**
Data sources are at the core of the gene prioritization problem since the quality of the predictions directly correlates with the quality of the data used to make these predictions. The different genomic data sources can be defined as different views on the same object, a gene. For instance, pathway databases (such as Reactome [58] and Kegg [59]) define a 'bio-molecular view' of the genes, while PPI networks (such as HPRD [60] and MINT [61]) define an 'interactome view'. A single data type might not be powerful enough to predict the disease causing genes accurately while the use of several complementary data sources allow much more accurate predictions [8, 29]. Supplementary Table 1 contains the list of the 12 data sources we have defined.

**Inputs**
Two distinct types of inputs can be distinguished: the prior knowledge about the genetic disorder of interest and the candidate search space. On the one hand, the prior knowledge represents what is currently known about the disease under study, it can be represented either as a set of genes known to play a role in the disease or as a set of keywords that describe the disease. On the other hand, the candidate search space defines which genes are candidates. For instance, a locus linked to a genomic condition defines a quantitative trait locus (QTL), the candidates are therefore the genes lying in that region. Another possibility is a list of genes differentially expressed in a tissue of interest that are not necessary from the same chromosomal location. Alternatively, the whole human genome can be used. An overview of the inputs required by the applications can be found in Table 2.

**Outputs**
For the 19 selected applications, the output is either a ranking of the candidate genes, the most promising genes being ranked at the top, or a selection of the most promising candidates, meaning that only the most promising genes are returned. Several tools are performing both at the same time (Gentrepid, Bitola, PosMed), that is first selecting the most promising candidates and then ranking only these. Several tools benefit from an additional output, a statistical measure, often a *P*-value, which estimates how likely it is to obtain that ranking by chance alone. The statistical measure is often of crucial importance since there will always be a gene ranked in first position even if none of the candidate genes is really interesting. Notice then that a selection can be obtained from a ranking by using the statistical measure (e.g. by choosing a threshold above which all the genes are considered as promising). You can find an overview of the outputs produced by the different applications in Table 2.

**Text mining**
It is the automatic extraction of information about genes, proteins and their functional relationships from text documents [62].

several gene prioritization methods, such as CAESAR [24], GeneRank [25] and CGI [26] propose interesting alternatives (e.g. natural language processing based disease model [24]), however, they only provide a standalone application to install and run locally. We believe that a web application is essential since it does not require an extensive IT knowledge to be installed and used. Second, there are methods that were once pioneers in that field and for which web applications were provided in the past, but are not accessible any more (e.g. TrAPSS [27], POCUS [28], Prioritizer [29]). Prioritizer recently moved from a living web application to a program to download and was therefore excluded prior to publication. Third, several studies also present case specific approaches tailored to answer a specific problem [30, 47–53]. For instance, Lombard *et al.* [47] have prioritized 10 000 candidates for the fetal alcohol syndrome (FAS) using a complex set of 29 filters. Their analysis reveals interesting therapeutic targets like TGF-$\beta$, MAPK and members of the Hedgehog signaling pathways. Another example is the network-based classification of breast cancer metastasis developed by Chuang *et al.* [48]. These approaches are, however, case specific and cannot be easily ported to another disease. Last, alternative techniques to circumvent recurrent problems in gene prioritization are currently under development. As an illustration, Nitsch *et al.* [31] have proposed a data-driven method in which knowledge about the disease under study comes from an expression data set instead of a training set or a keyword set.

## DESCRIPTION OF THE GENE PRIORITIZATION METHODS

### The genomic data are at the core

We have defined a data source as a type of data that represents a particular view of the genes (see Box 1—'Gene view') and thus can correspond to several

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

related databases. Data sources are at the core of the gene prioritization problem since both high coverage and high quality data sources are needed to make accurate predictions. In total, we have defined 12 data sources: text mining (co-occurrence and functional mining), protein–protein interactions, functional annotations, pathways, expression, sequence, phenotype, conservation, regulation, disease probabilities and chemical components. Using these categories, we have built a data source landscape, which describes for each tool which data sources it uses (Supplementary Table 1). We can observe from the data source landscape map that text mining is by far the most widely used data source since 14 out of the 19 tools are using co-occurrence or functional text mining. Most of the approaches make use of a wide range of data sources covering distinct views of the genes, but four tools rely exclusively on text mining (PGMapper, Bitola, aGeneApart and GeneProspector), however their use of advanced text mining techniques still allow them to make novel predictions. At the other end of the spectrum, conservation, regulation, disease probabilities and chemical components are poorly used and only by two tools at most although they describe unique features that might not always be captured by the other data sources. However, the rule should not be to include as many data sources as possible but rather to reach a critical mass of data beyond which accurate predictions can be made.

## Inputs and outputs of the methods

The tools also differ in the inputs they require and the outputs they provide. Two types of inputs have been distinguished: the prior knowledge about the genetic disorder of interest and the candidate search space. We furthermore consider two possibilities for the prior knowledge as it can be defined by a set of genes or by a set of keywords. The retrieval of a training set requires the knowledge of, at least, one disease causing gene, but preferably more than one. In addition, the set needs to be homogeneous, meaning that it usually contains between 5 and 25 genes that, together, describe a specific biological process. When no disease gene can be found, members of the pathways disturbed by the diseases are also an option (Thienpont *et al.*, manuscript in preparation). Alternatively, several tools accept text as input, text is either a disease name, selected from a list, or a set of user defined keywords that describe the disease under study. In the second case, the expert should define a complete set of keywords that covers most aspects of the disease (e.g. to obtain reliable results, 'diabetes' should be used in conjunction with 'insulin', 'islets', 'glucose' and others diabetes related keywords but not alone). Regarding the candidate search space, we have distinguished between a locus, a differentially expressed genes (DEG) list, and the whole genome. A locus is a set of neighboring genes (e.g. all genes from the cytogenetic band 22q11.23) while the genes in a DEG list are not necessarily located at the same locus. Although these two options are similar, the distinction we made is important since several tools allow the definition of a locus but not of DEG list and vice versa. Alternatively, nine tools allow the exploration of the full genome, in case no candidate gene set can be defined beforehand.

Regarding the outputs, two types were considered, a ranking and a selection of the candidate genes. In a ranking scenario, all the candidates are ranked so that the most promising candidate can be found at the top, while for a selection, a subset of the original candidate set, containing only the most promising candidates, is returned. From the 19 tools, four perform a selection of the candidates and three of these four perform a selection followed by a ranking. In addition, we record which tools further measure the significance of their results via any statistical method. Of interest, a selection can then be obtained from a ranking by using a threshold on this statistical measure. Table 2 shows an overview of the input data required by the tools as well as the output they produce. Also, a clustering of the tools regarding to their inputs and outputs is presented in Figure 3. In addition, we have created a decision tree to help users to choose the most suitable tools for their biological question. The tree is based on three basic questions that users should ask themselves before selecting the tools they want to use. By answering these questions, users define first, which genes are candidate; second, how the current knowledge is represented; and third (when necessary), what is the desired output type.

## The importance of biological validation

Since the methods we are interested in are predictive, an important criterion for selection is the performance. The tools reviewed here were all originally published together with the results of a benchmark analysis as a proof of concept. It is however difficult to

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

**Table 2:** Description of the inputs needed by the tools and the outputs produced by the tools

| Tool | Inputs | | | | | Output | | |
|---|---|---|---|---|---|---|---|---|
| | Training data | | Candidate genes | | | Ranking | Selection of candidates | Test statistic |
| | KnownGenes | Keywords | Region | DEG | Genome | | | |
| SUSPECT | | x | x | x | | x | | |
| ToppGene | x | | | x | | x | | x |
| PolySearch | | x | | | x | x | | x |
| MimMiner | | x | | | x | x | | |
| PhenoPred | | x | | | x | x | | |
| PGMapper | | x | x | x | | x | | |
| Endeavour | x | x | x | x | x | x | | x |
| G2D | x | x | x | | | x | | x |
| TOM | x | | x | | | | x | |
| SNPs3D | | x | | | x | x | | |
| GenTrepid | | x | x | x | | x | x | |
| GeneWanderer | x | | x | x | | x | | x |
| Bitola | | x | x | | x | x | x | |
| CANDID | | x | | | | x | | x |
| aGeneApart | | x | | | x | x | | x |
| GeneProspector | | x | | | x | x | | |
| PosMed | | x | x | | x | x | x | x |
| GeneDistiller | x | x | x | | | x | | |

We distinct two types of inputs: the prior knowledge about the genetic disorder of interest and the candidate search space. The prior knowledge can be represented either as a set of genes known to play a role in the disease or as a set of keywords that describe the disease. The candidate search space is either a locus linked to a genomic condition or a list of genes differentially expressed in a tissue of interest (DEG) or the whole human genome. The output is either a ranking of the candidate genes or a selection of the most promising candidates. In addition, a statistical measure that estimates how likely it is to obtain that result by chance alone. More details about the inputs and outputs can be found in the Box I.

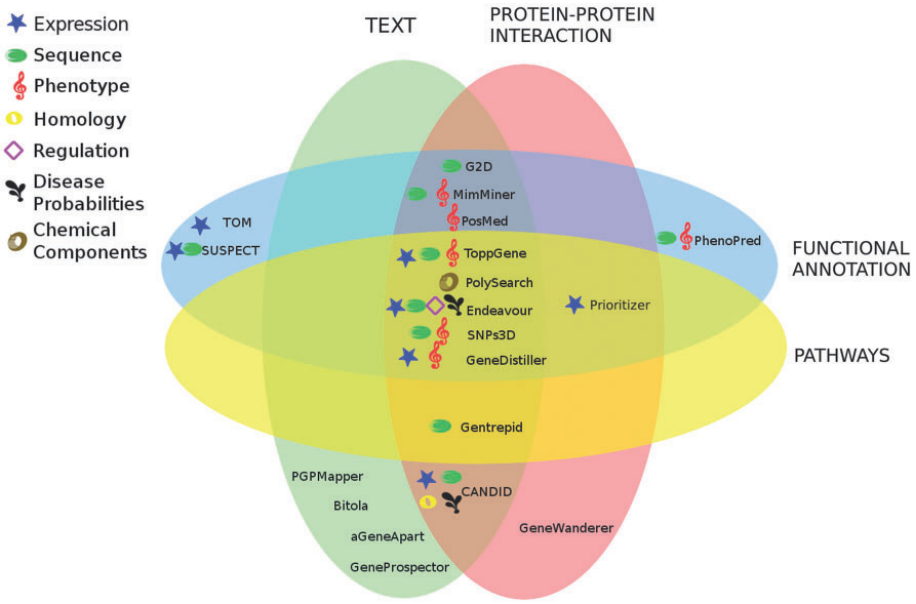



**Figure 3:** Repartition of the 19 tools according to the data sources they use. The four data sources most commonly used are Text (functional and interactions mining), protein–protein interactions, functional annotations and pathways and are therefore represented as large ellipses. The additional seven data sources are represented with symbols.

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

compare the performance of these benchmarks directly since their setups are different (different diseases, different genes). Although a rigorous comparison is still missing, various studies that compare several gene prioritization tools by analyzing their performance on a particular disease have been performed (e.g. on T2D [41, 42, 54]). An overview is presented in Supplementary Table 2. Although it is of primary importance, the performance obtained through a benchmark analysis represents more a proof of concept than a critical performance assessment. Therefore, it is only an estimation of the real performance (e.g. for a real biological application) and it is also most likely benchmark specific. That is the reason why we believe that the definition of the desired inputs/outputs and data sources, and the knowledge of real biological applications are also crucial.

Beside these benchmarks, several biological applications have been described in the literature. Supplementary Table 3 gives an overview of these applications. Interestingly, three of them analyzed T2D associated loci and are using several gene prioritization tools in conjunction [41, 42, 54]. Elbers *et al.* [42] analyzed five loci previously reported to be linked with both T2D and obesity that encompass more than 600 genes in total. The authors used six gene prioritization tools in conjunction and reported 27 interesting candidates. Some of them were already known to be involved in either diabetes or obesity (e.g. TCF1 and HNF4A, responsible for maturity onset diabetes of the young, MODY) but some candidates were novel predictions. Among them, five genes were involved in immunity and defense (e.g. TLR2, FGB) and it is known that low-grade inflammation in the visceral fat of obese individuals causes insulin resistance and subsequently T2D. Also, 10 candidate genes were so-called 'thrifty genes' because of their involvement in metabolism, sloth and gluttony (e.g. AACS, PTGIS and the neuropeptide Y receptor family members). Using a similar strategy, Tiffin *et al.* [41] prioritized T2D and obesity associated loci and proposed another set of 164 promising candidates. Of interest, 4 of the 27 candidates reported by Elbers *et al.* were also reported by Tiffin *et al.* (namely CPE, LAMA5, PPGB and PTGIS). Although there is an overlap between the predictions, some important discrepancies remain and can be explained by the fact that the two studies do not focus on the same set of loci and do not use the same gene prioritization tools. This indicates that several gene prioritization tools can be applied in parallel to strengthen the results. Teber *et al.* [54] compared the finding from recent genome-wide association studies (GWAS) to the predictions made by eight gene prioritization methods. Of the 11 genes associated with highly significant SNPs identified by the GWAS, eight were flagged as promising candidates by at least one of the method. Another interesting validation is a computationally supported genetic screen performed by Aerts *et al.* [43] in fruit fly. The aim of a genetic screen is to discover *in vivo* associations between genotypes and phenotypes. A forward genetic screen is usually performed in two steps: in the first step, the loci associated to the phenotype under study are identified and in a second step, the genes from these loci are assayed individually. Aerts *et al.* have introduced a computationally supported genetic screen in which the associated loci found in the first step are prioritized using Endeavour and then only the genes ranked in the top 30% of every locus are assayed in a secondary screen. Additionally, it was shown that 30% is a conservative threshold since all the positives were ranked in the top 15%. This shows that gene prioritization tools, when integrated into such workflows, can increase their efficiency for a decreased cost.

## Intuitive interfaces

Beside the data, the inputs/outputs and the performance, what is critical for a tool to be used is its interface. Ideally, it has to be an intuitive interface that accepts simple inputs and provides detailed outputs. A past success and reference in bioinformatics is basic local alignment search tool (Blast) for which only a single sequence needs to be provided [55]. In return, Blast provides the complete detailed alignments together with cross-links to sequence databases so that the user can fully understand why the input sequence matches to a given database sequence. We, as a community, should develop tools that answer the end users' needs and that probably corresponding to the simple input—detailed output paradigm described above. Besides, the presence of an advanced mode that allows users to fine tune the analysis is also clearly an advantage (e.g. defining a threshold for the Blast *e*-value).

Several gene prioritization tools such as MimMiner, PhenoPred, aGeneApart and GeneProspector can already be fed with a single

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

disease name that represents the simplest training input possible. However, an advanced mode to fine tune the analysis is missing for these applications. The outputs generated by the tools are very detailed and almost always contain cross-references to external databases (e.g. Hugo, EnsEMBL, RefSeq). However, only few tools present detailed information about the data underlying the ranking of the candidate genes. This data is crucial for the user who needs to determine which candidates should be investigated further. This is probably the weakest point of most of the current tools although several tools like Suspects and G2D already propose preliminary solutions. In addition, most of the tools benefit from a user manual and a dedicated help section that help users to understand how they should interact with the interface.

## FUTURE DIRECTIONS

With the use of advanced high-throughput technologies, the amount of genomic data is growing exponentially and the quality of the gene prioritization methods is also increasing accordingly. However, several avenues need to be explored in the coming years to increase even further the potential of these tools. We have already mentioned the interface, which is sometimes overlooked in the software development process. More at the data level, some efforts have already been made to use the huge amount of data available for species close to human [9–12]. Already, several tools described in the current review include rodent data (e.g. SNPs3D, ToppGene, GeneDistiller, Posmed). However, the development of gene prioritization approaches combining in parallel many data sources from different organisms is still to come. Another important development is the inclusion of clinical and patient related data. DECIPHER [56] already represents a first step in that direction since it includes aCGH data from patients and allow text mining prioritization (using the core engine of aGeneApart [18]) of the genomic alterations, detected in the aCGH data, with respect to the phenotype of the patient. Efforts should also be made to include data sources that have been, so far, rarely included such as chemical components and miRNA data. Another important research track is to explore different computational approaches to improve once more the algorithms that are running the gene prioritization methods. Preliminary results have shown, for example, that kernel methods are more efficient than simpler statistical methods such as Pearson correlation or binomial based over-representation [57]. The last challenge of this field is its necessary adaptation to the shift observed in genetics towards the study of more complex disorders that is though to be more difficult than the study of the Mendelian diseases.

Altogether, the methods described in this review represent significant advances indicating that this field is still an emerging field. It is therefore most likely that novel methods will be developed in the future and that the existing ones will be improved. To overcome the limitations due to the static nature of this review, we have developed a website whose aim is to represent an updatable electronic version of the present review. This web site, termed 'Gene Prioritization Portal' (available at: http://www.esat.kuleuven.be/gpp), contains, for every tool, a detailed sheet that summarizes the necessary information such as the inputs needed and the data sources used. It also builds tables that describe the general data source usage and the general input/output usage that are equivalent to Table 2 and Supplementary Table 1 of the current publication. We believe that this website represents a first step to guide users through their gene prioritization experiments.

## CONCLUSION

This review tries to clarify the world of gene prioritization to the final user through an exhaustive guide of 19 human candidate gene prioritization methods that are freely accessible through a web interface. This taxonomy has been done according to different characteristics of the tools, including the type of input, data sources used during the process of prioritization and the desired output. We think that this review is a useful tool not only to help the wet lab researchers to dive into gene prioritization, but also to guide them to select the most convenient method for their analysis.

To keep up with the especially fast evolving world of bioinformatics in general and gene prioritization in particular, we have developed a website http://www.esat.kuleuven.be/gpp/ that contains updated information of all the tools described in this review. We expect our portal to become a reference point in gene prioritization where not only users but also developers will find up-to-date information necessary for their research.

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

**Key Points**

- Numerous computational methods have been developed to tackle the gene prioritization problem in human; we have collected the methods that offer such web services freely.
- We have described how these methods differ from each other by the inputs they need, the outputs they produce and the data sources they use.
- We have furthermore described some of the biological applications to which gene prioritization approaches were successfully applied.
- A website that contains information about the available gene prioritization methods has been developed and will be updated on a regular basis.

## SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

## FUNDING

Research Council KUL [GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA]; the Flemish Government [G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, ICCoS, ANMMM, MLDM, G.0733.09, G.082409, GBOU-McKnow-E, GBOU-ANA, TAD-BioScope-IT, Silicos, SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM-IOTA3, O&O-Dsquare]; the Belgian Federal Science Policy Office [IUAP P6/25]; and the European Research Network on System Identification (ERNSI) [FP6-NoE, FP6-IP, FP6-MC-EST, FP6-STREP, FP7-HEALTH].

### *References*

1. Redon R, Ishikawa S, Fitch KR, *et al*. Global variation in copy number in the human genome. *Nature* 2006;**444**:444–54.
2. Marazita ML, Murray JC, Lidral AC, *et al*. Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35. *Am J Hum Genet* 2004;**75**:161–73.
3. Jorde LB. Linkage disequilibrium and the search for complex disease genes. *Genome Res* 2000;**10**:1435–44.
4. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. *Nat Genet* 2002;**31**:316–9.
5. Zhang P, Zhang J, Sheng H, *et al*. Gene functional similarity search tool (GFSST). *BMC Bioinformatics* 2006;**7**:135.
6. Adie EA, Adams RR, Evans KL, *et al*. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 2006;**22**:773–4.
7. Hutz JE, Kraja AT, McLeod HL, *et al*. CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet Epidemiol* 2008;**32**:779–90.
8. Aerts S, Lambrechts D, Maity S, *et al*. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;**24**:537–44.
9. Chen J, Xu H, Aronow BJ, *et al*. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* 2007;**8**:392.
10. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 2006;**7**:166.
11. Seelow D, Schwarz JM, Schuelke M. GeneDistiller–distilling candidate genes from linkage intervals. *PLoS ONE* 2008;**3**:e3874.
12. Yoshida Y, Makita Y, Heida N, *et al*. PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res* 2009;**37**:W147–52.
13. Rossi S, Masotti D, Nardini C, *et al*. TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res* 2006;**34**:W285–92.
14. Masotti D, Nardini C, Rossi S, *et al*. TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders. *Bioinformatics* 2008;**24**:428–9.
15. Cheng D, Knox C, Young N, *et al*. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res* 2008;**36**:W399–405.
16. van Driel MA, Bruggeman J, Vriend G, *et al*. A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**:535–42.
17. Hristovski D, Peterlin B, Mitchell JA, *et al*. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* 2005;**74**:289–98.
18. Van Vooren S, Thienpont B, Menten B, *et al*. Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res* 2007;**35**:2533–43.
19. Yu W, Wulf A, Liu T, *et al*. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics* 2008;**9**:528.
20. George RA, Liu JY, Feng LL, *et al*. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* 2006;**34**:e130.
21. Xiong Q, Qiu Y Gu W. PGMapper: a web-based tool linking phenotype to genes. *Bioinformatics* 2008;**24**:1011–3.
22. Köhler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**:949–58.
23. Radivojac P, Peng K, Clark WT, *et al*. An integrated approach to inferring gene-disease associations in humans. *Proteins* 2008;**72**:1030–7.
24. Gaulton KJ, Mohlke KL, Vision TJ. A computational system to select candidate genes for complex human traits. *Bioinformatics* 2007;**23**:1132–40.
25. Morrison JL, Breitling R, Higham DJ, *et al*. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* 2005;**6**:233.

Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

26. Ma X, Lee H, Wang L, *et al*. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007;**23**: 215–21.
27. Braun TA, Shankar SP, Davis S, *et al*. Prioritizing regions of candidate genes for efficient mutation screening. *Hum Mutat* 2006;**27**:195–200.
28. Turner FS, Clutterbuck DR, Semple CAM. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 2003;**4**:R75.
29. Franke L, van Bakel H, Fokkens L, *et al*. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;**78**:1011–25.
30. Tiffin N, Okpechi I, Perez-Iratxeta C, *et al*. Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes. *Physiol Genomics* 2008;**35**:55–64.
31. Nitsch D, Tranchevent L, Thienpont B, *et al*. Network analysis of differential expression for the identification of disease-causing genes. *PLoS ONE* 2009;**4**:e5526.
32. Tranchevent L, Barriot R, Yu S, *et al*. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res* 2008;**36**:W377–84.
33. Perez-Iratxeta C, Wjst M, Bork P, *et al*. G2D: a tool for mining genes associated with disease. *BMC Genet* 2005;**6**: 45.
34. Perez-Iratxeta C, Bork P, Andrade-Navarro MA. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res* 2007;**35**:W212–6.
35. Smith NGC, Eyre-Walker A. Human disease genes: patterns and predictions. *Gene* 2003;**318**:169–75.
36. Goh K, Cusick ME, Valle D, *et al*. The human disease network. *Proc Natl Acad Sci USA* 2007;**104**:8685–90.
37. Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. *Nature* 2001;**409**:853–5.
38. Iizuka M, Kubo Y, Tsunenari I, *et al*. Functional characterization and localization of a cardiac-type inwardly rectifying K+ channel. *Recept Channels* 1995;**3**:299–315.
39. Wasada T. Adenosine triphosphate-sensitive potassium (K(ATP)) channel activity is coupled with insulin resistance in obesity and type 2 diabetes mellitus. *Intern Med* 2002;**41**: 84–90.
40. Lavine N, Ethier N, Oak JN, *et al*. G protein-coupled receptors form stable complexes with inwardly rectifying potassium channels and adenylyl cyclase. *J Biol Chem* 2002; **277**:46010–19.
41. Tiffin N, Adie E, Turner F, *et al*. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res* 2006; **34**:3067–81.
42. Elbers CC, Onland-Moret NC, Franke L, *et al*. A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol Metab* 2007;**18**:19–26.
43. Aerts S, Vilain S, Hu S, *et al*. Integrating computational biology and forward genetics in Drosophila. *PLoS Genet* 2009;**5**:e1000351.
44. Myers CL, Barrett DR, Hibbs MA, *et al*. Finding function: evaluation methods for functional genomic data. *BMC Genomics* 2006;**7**:187.
45. Troyanskaya OG. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinformatics* 2005;**6**:34–43.
46. Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput Biol* 2008;**4**: e1000160.
47. Lombard Z, Tiffin N, Hofmann O, *et al*. Computational selection and prioritization of candidate genes for fetal alcohol syndrome. *BMC Genomics* 2007;**8**:389.
48. Chuang H, Lee E, Liu Y, *et al*. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007;**3**:140.
49. Huang Q, Li GHY, Cheung WMW, *et al*. Prediction of osteoporosis candidate genes by computational disease-gene identification strategy. *J Hum Genet* 2008;**53**:644–55.
50. Gajendran VK, Lin J Fyhrie DP. An application of bioinformatics and text mining to the discovery of novel genes related to bone biology. *Bone* 2007;**40**:1378–88.
51. Alsaber R, Tabone CJ, Kandpal RP. Predicting candidate genes for human deafness disorders: a bioinformatics approach. *BMC Genomics* 2006;**7**:180.
52. Rasche A, Al-Hasani H, Herwig R. Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. *BMC Genomics* 2008;**9**:310.
53. Furney SJ, Calvo B, Larrañaga P, *et al*. Prioritization of candidate cancer genes–an aid to oncogenomic studies. *Nucleic Acids Res* 2008;**36**:e115.
54. Teber ET, Liu JY, Ballouz S, *et al*. Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies. *BMC Bioinformatics* 2009;**10**(Suppl. 1):S69.
55. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
56. Firth HV, Richards SM, Bevan AP, *et al*. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* 2009; **84**:524–33.
57. De Bie T, Tranchevent L, van Oeffelen LMM, *et al*. Kernel-based data fusion for gene prioritization. *Bioinformatics* 2007; **23**:i125–32.
58. Vastrik I, D'Eustachio P, Schmidt E, *et al*. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 2007;**8**:R39.
59. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
60. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human Protein Reference Database–2009 update. *Nucleic Acids Res* 2009;**37**:D767–72.
61. Chatr-aryamontri A, Ceol A, Palazzi LM, *et al*. MINT: the Molecular INTeraction database. *Nucleic Acids Res* 2007;**35**: D572–4.
62. Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. *Genome Biol* 2005; **6**:224.


Downloaded from http://bib.oxfordjournals.org at KU Leuven on March 22, 2010

# Chapter 3

# *Study 2*: Gene prioritization - a critical assessment

Beyond the conceptual differences among gene prioritization tools as described in *Study 1*, the performance in actually identifying novel disease genes from long lists of candidate genes is an essential issue for choosing a gene prioritization tool. In this study, we aim to compare gene prioritization tools that are available as web applications in terms of their performance. For critically assessing the usefulness and practicability of gene prioritization tools we have designed a workflow in which we have selected recently reported disease gene associations from literature to make predictions. By doing this, we mimic novel discovery which estimates a more accurate performance for tools comparison.

**Publication**
This work is still in preparation and will be submitted in the near future.
*Reference:* Nitsch D*, Tranchevent L*, Capdevila FB*, Devriendt K, De Moor B, De Causmaecker P, Moreau Y (2011). A Critical Assessment of Gene Prioritization. Internal Report, 11-15, ESAT-SISTA, K.U.Leuven (Leuven, Belgium). *Equally contributed

**Personal contribution**
This work represents a collaborative effort with equal contribution from the PhD candidate and two co-authors in terms of developing the idea, setting up the study, performing the experiments, analyzing the data and writing the paper.

# CAGP – A Critical Assessment of Candidate Gene Prioritization

**Daniela Nitsch[1,*], Léon-Charles Tranchevent[1*], Francisco Bonachela Capdevila[2*], Koenraad Devriendt[3], Bart de Moor[1], Patrick De Causmaecker[2], Yves Moreau[1§]**

[1] Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium
[2] CODeS Group, Department of Computer Science, Katholieke Universiteit Leuven, Campus Kortrijk, Belgium
[3] Center for Human Genetics, University Hospitals Leuven, Leuven, Belgium
*equally contributed

## ABSTRACT

*During the last years, distinct gene prioritization methods have been proposed to rank large sets of candidate genes. Some of these methods have been implemented into publicly available web applications that users can freely access and use. In this study, we compare the performance of several freely accessible web based gene prioritization tools on novel data, which, to our knowledge, has never been performed before. We have designed a workflow in which we select recently reported disease gene associations from literature and use several gene prioritization tools to make predictions. Our approach mimics a novel discovery, therefore the estimation of the performance is more accurate and tools can be compared.*

## INTRODUCTION

A major challenge in human genetics is to discover novel Mendelian disorders causing genes or genes associated to complex traits. This is a crucial process because understanding the disease mechanisms is the first step towards the development of an effective treatment. The discovery of a novel disease gene often starts with a cytogenetic study, a linkage analysis, an high-throughput 'omics' experiment or a genome wide association studies (GWAS). However, these studies do not always reveal the exact disease mechanisms but often result in large lists of candidate genes that are potentially relevant (Hardy *et al.*, 2009). Similarly, the recent advances in new generation sequencing offer promising opportunities to explore the genome alterations of patients (Schuster *et* al, 2008). However, thousands of mutations in hundreds of genes are often detected, among which only a few are in fact linked to the genetic condition of interest (Lupski *et al.,* 2010). The experimental validation of these candidate genes, for instance through sequencing, pathway or expression analysis, is still very expensive and time consuming. An efficient way to reduce the validation costs is to narrow down the large list of candidate genes to a small and manageable set of highly promising genes, a process termed gene prioritization. Prioritization was in the past achieved manually by geneticists and biologists and was mainly based on their own expertise. Nowadays, biologists and geneticists can use computational approaches that can handle and analyze the large amount of genomic data available to them in the post-genomic era.

In the last decade, many gene prioritization methods have been proposed, some of which have been implemented into publicly available applications that users can freely access and use. We have recently reviewed the web based gene prioritization applications and described how they differ by the inputs they require, the outputs they produce and the data they use (Tranchevent et al., 2010). This information is summarized in our Gene Prioritization Portal (http://www.esat.kuleuven.be/gpp) that currently describes 24 prioritization tools. This website has been designed to help researchers to carefully select the tools that best correspond to their needs. For instance, only few tools can prioritize the whole genome, which can be necessary when

no positive regions can be identified beforehand. Another example is the study of a poorly characterized disorder, for which a prioritization method that do not rely on a set of known disease genes might be required.

However, beyond these conceptual differences, one essential parameter to consider when selecting gene prioritization methods is their performance: i.e., their ability to identify novel disease genes among large lists of candidate genes. A common standard in bioinformatics is to estimate the performance with a benchmark analysis. Several publications that introduce a novel prioritization approach also describe a comparative benchmark with several already existing tools (Hutz et al., 2008, Köhler et al., 2008, Thornblade et al., 2007). However, these benchmarks are most of the time cross-validations of gold-standard disease datasets (e.g., known data). And therefore, the estimation of the performance is likely an overestimate of the real performance (i.e., on novel data) (Myers *et al.,* 2006). There are studies that combine the predictions made by several prioritization tools when analyzing disease associated loci, mostly for type 2 diabetes and obesity (Elbers *et al.,* 2007; Teber *et al.,* 2009; Tiffin *et al.,* 2006). However, the results have not been experimentally validated meaning that it is not possible to identify which methods made the more accurate predictions. There exist a few studies that combine computational and experimental analysis, i.e., in silico generated hypothesis are then validated in vivo. For instance, we have developed a computationally supported genetic screen in drosophila that lead to the identification of 12 novel atonal genetic interactors (Aerts et al., 2009). Although very useful, these studies often rely on the use of a single tool and can't therefore be used to compare different approaches.

In this study, we aim at comparing the performance of several freely accessible web based gene prioritization tools on novel data, which, to our knowledge, has never been performed before. To this aim, we designed a workflow that resembles the Critical Assessment of protein Structure Prediction (CASP) concept. In the case of CASP, a protein 3D structure is known but kept secret while dedicated algorithms predict the structure of that protein. These predictions are then compared to the real structure, and algorithms can therefore be compared. In our case, we select recently reported disease gene associations from literature and use several gene prioritization tools to make predictions. Our approach relies on the fact that, when the prioritization tools are used, the novel disease gene association of interest is not yet included in the databases that underlie these tools. As a consequence, our approach mimics a novel discovery, therefore the estimation of the performance is more accurate and tools can be compared.

## MATERIAL AND METHODS

### Gene prioritization tools

We aim at comparing the gene prioritization tools that can easily be used, and therefore only select the tools for which a free web based implementation is available. The main objective is to assess the ability of the gene prioritization tools to discover novel disease gene. We therefore don't select the tools whose ranking strategies depend exclusively on text mining (Van Vooren *et al.,* 2007; Hristovski *et al.,* 2005; Yu *et al.,* 2008; Xiong *et al.,* 2008). In total, we select seven tools: SUSPECTS (Adie et al., 2006), ToppGene (Chen et al., 2007), GeneDistiller (Seelow et al., 2008), GeneWanderer (Köhler et al., 2008), Posmed (Yoshida et al., 2009), Candid (Hutz et al., 2008), and Endeavour (Aerts et al., 2006). For an overview of the tools, please see Supplementary Table S1. All tools except Candid are used to prioritize a set of candidate genes (from a chromosomal region), and Candid is used to prioritize the whole genome. Endeavour offers both genome-wide and candidate set based prioritizations and is therefore used for both (Endeavour-GW for genome-wide prioritization, Endeavour-CS for the candidate set prioritization). In addition, GeneWanderer can be run with up to four different ranking strategies (random walk, diffusion kernel, shortest path and direct interaction). We present the results for the first two strategies (GeneWanderer-RW for random walk, GeneWanderer-DK for diffusion kernel) since they have been showed to outperform the other two, simpler, methods and since they can be efficiently used with many training genes.

Posmed performance shows a strong dependency on the inputted keywords and we run it twice with different inputs. In the first run, we use the complete keyword set (Posmed-KS), and in the second, we only use the name of the disease (Posmed-DN). GeneDistiller is trained with both genes and keywords. These keywords are then used to find additional genes through the mining of OMIM, which in our case has less influence since OMIM was already used to derive the training genes. We therefore consider that GeneDistiller is trained with genes only. Candid is the only tool that can also be trained with disease specific tissues, when available, tissues relevant to the disease under study are used. Except when mentioned, the tools are run with the default options, in particular no fine tuning of the parameter is done. Of interest, SUSPECTS went offline during our study after the 28th association and is not supported anymore (Euan Adie, personal communication), therefore, SUSPECTS results are based on 28 associations over 42.

**Validation dataset**

The validation dataset is built by mining the scientific literature to identify the recently discovered disease-gene associations. This is achieved manually in order to avoid false positive associations. We select 6 journals that likely publish papers that describe such associations: Nature Genetics, American Journal of Medical Genetics (part A / part B), Human Genetics, Human Molecular Genetics, and Human Mutation. We select all the novel disease-gene associations regardless of the disease under study, of the methodology used, and of whether the findings were confirmed or not. Novelty is assessed by using OMIM (Mc Kusick et al., 1998), the Genetic Association Database (Becker et al., 2004), GoPubmed (www.gopubmed.org) and, GeneCards (www.genecards.org). This process is kept active for 6 months (May 15 – November 15, 2010) and leads to a collection of 42 associations (please see Supplementary Table S2 for an overview). For each association, the tools are run as soon as the association is identified following the defined workflow. By doing this, we simulate as much as possible the prediction of a novel disease gene since the underlying databases are still unaware of the association.

Once an association is identified, the exact inputs for the different tools have to be defined. For instance, ToppGene, GeneDistiller, GeneWanderer and Endeavour require training genes (genes already known to be associated to the disease under study) whereas SUSPECTS, Posmed, GeneDistiller and Candid require keywords that describe the disease. Training genes and keywords were collected from the corresponding OMIM pages, GAD pages and possibly from recent reviews (from PubMed). BioMart (Haider et al., 2009) is used to map between gene symbols and tool specific gene identifiers (e.g., EntrezGene or Ensembl identifiers). As mentioned above, most of the tools require in addition a set of candidate genes (from the whole genome). Several tools accept chromosomal coordinates whereas some prefer cytogenetics bands. For each association, we select the cytogenetics bands that cover approximately 10Mb around the novel disease gene and derive the chromosomal coordinates. We choose 10Mb in order to obtain on average at least 100 candidate genes. Once again, BioMart is used to retrieve specific gene identifiers. For an overview of the inputs for the 42 associations, please see Supplementary Table S3.

**Statistical measures**

For each tool, we then assess its ability to discover novel disease genes with several statistical measures. We first compute the median of the rank ratio over all associations. We preferably use rank ratio over rank because tools do not necessarily return the same number of candidate genes even when fed with the same inputs. In addition, we also draw the boxplots to give a more comprehensive view of the tools performance. An efficient method to compare the tools is to build the Receiver Operating Characteristic (ROC) curves, and to compute the Area Under the Curve (AUC) as an estimate of the global performance. Notice that, in our setup, the AUC is equivalent to one minus the average rank ratio. To compare tools further, we computed the sensitivity at 10% and at 30%. This is motivated by the fact that in a real situation, the number of candidate genes to assay often needs to be limited due to financial and time constraints. We have selected two thresholds that represent reasonable biological hypothesis as we illustrated in a genetic screen

(Aerts et al., 2009). The corresponding sensitivity measures are used to estimate how efficient the tools are if only the top 10% or 30% candidate genes would be assayed.
There are cases for which some tools are not able to find the novel disease gene at all, we therefore include a reliability measure, simply the percentage of associations in which each tool is able to prioritize the novel disease gene. Lastly, we also compute a heat map to detect any correlation between tools.

## RESULTS

The overall ranking results of all gene prioritization tools are presented in Table T1 and Table T2.

**Rank Ratio**
If we compute the median of the rank ratios, GeneDistiller, Endeavour-CS, and SUSPECTS are the tools that perform best (respectively 11.2%, 11.2%, and 11.4%). They are followed by Endeavour-GW (15.9%), Candid (17.9%), ToppGene (20.3%), GeneWanderer-RW (22%), and GeneWanderer-DK (23.3%), Posmed-KS (29.5%, and Posmed-DN (37.7%). For more details see Table T1 and Table T2.
The boxplotes presented in Figure 1 illustrate that both, GeneDistiller and Endeavour perform better than the other candidate set based prioritization tools (Figure 1-A). Among the genome-wide tools, Endeavour performs better than Candid (Figure 1-B).

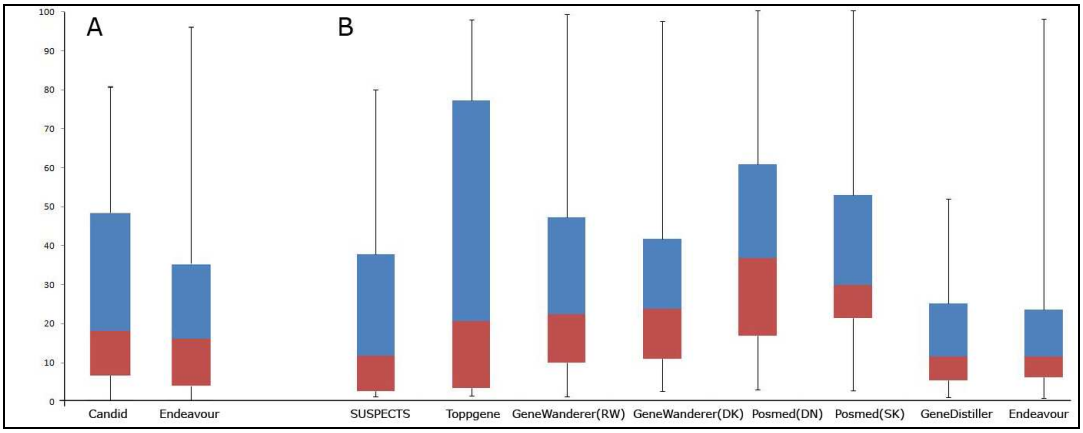



**Figure 1: Ranking results of the novel disease genes from the validation dataset illustrated as boxplots for the genome-wide (A) and candidate gene set based (B) prioritization tools.**

**Reliability**
When considering reliability, Endeavour and Candid performed the best with 100% closely followed by ToppGene, GeneWanderer, and GeneDistiller with more than 95% (meaning that only one or two associations are not found). At the other hand of the spectrum, Posmed -KS and Posmed-DN only works for about half of the experiments (respectively 48.8% and 51.1%). Supplementary Figure S1 illustrates the reliability of all tools. For more details see Table T1 and Table T2.

**Area Under the Curve**

When we compare the tools based on the global AUC (see Figure 2), we observe that GeneDistiller still appears as the best performing tool overall with an AUC of 86%. It is followed by Endeavour-CS (82%), Endeavour-GW (79%), SUSPECTS (77%), Candid (74%), GeneWanderer-RW (72%), GeneWanderer-DK (67%), Posmed-KS (67%), ToppGene (64%), and Posmed-DN (61%). The ROC curves are in general intertwined meaning that none of the approaches is clearly performing better than the other. However, se postulate that, in our case, the most important section of the ROC curve is the beginning and therefore introduce the sensitivity at 10% and at 30%. For these two measures, we observe very similar results. Indeed, at 10%, GeneDistiller is first with 20 associations found, followed by both SUSPECTS and Endeavour-CS with 18 associations for a random expectation of 4.3. At 30%, the best tool is Endeavour with 39 associations, followed by GeneDistiller with 34 associations found (random expectation is 13).

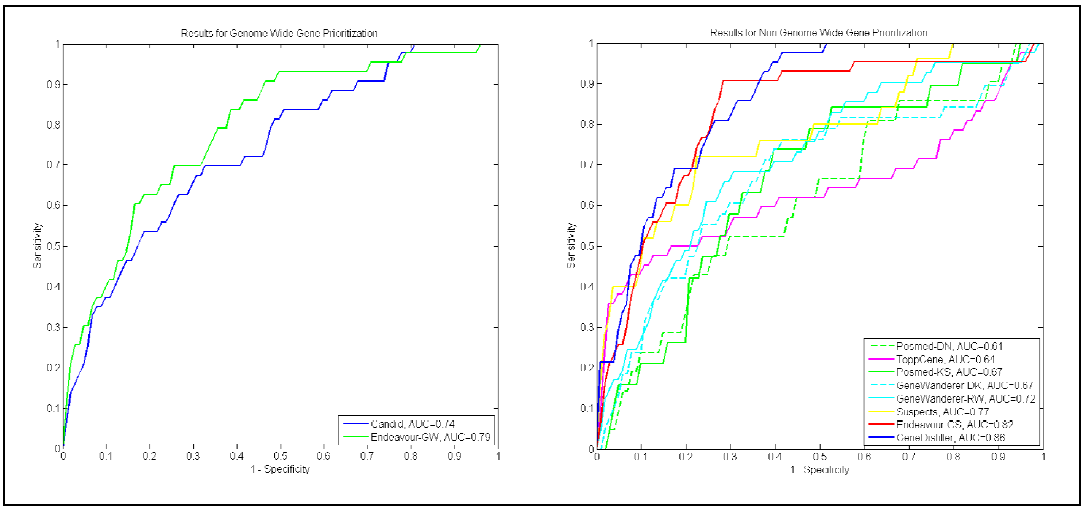



**Figure 2: ROC curves of the genome-wide (left) and candidate gene set based (right) prioritization tools.**

**Sensitivity**

Considering the sensitivity at 10% and 30% (see Table T1 and Table T2), we can confirm our findings: GeneDistiller shows the highest sensitivity at 10% with 20 and at 30% with 34 disease genes, whereas Endeavour and ToppGene are only slighty worse with 18 disease genes at 10%. However, Endeavour shows highest sensivity at 30% with 39 disease genes. The other tools show smaller sensitivity at both levels: SUSPECT (10/18), GeneWanderer-RW (11/27), GeneWanderer-DK (9/22), Posmed-KS (5/100, and Posmed-DN (3/11).

**Correlations**

Supplementary Figure S2 shows the heat map of the novel disease gene ranking positions for all tools in this study. For the tools that have two modes (i.e., Posmed, GeneWanderer, Endeavour), the two modes are highly correlated (at least 0.9). There is also a significant correlation between Candid and GeneWanderer (~0.8). The other values are within 0.4 and 0.7, indicating that all tools are moderately correlated.

## DISCUSSION

We aimed at assessing the usefulness of seven gene prioritization tools that were freely available via web applications. We have built a validation scheme that resembles the CASP scheme for protein structure prediction. The validation is based on 42 recently discovered disease-gene associations from literature and contains novel genes for both monogenic conditions and more complex disorders. We selected all the novel associations regardless of their strength, and of the underlying methodology. To mimic a real discovery, we have run the tools as soon as the article appeared online in order to use databases still unaware of the novel disease-gene association. This also means that we had to exclude tools that query PubMed online since their results would be biased. One exception is Candid that uses other data sources beside PubMed and for which prioritization without Pubmed can be achieved.

We want to compare the performance of the tools even if the inputs are different (training genes versus keywords, genome-wide versus candidate set). Among the seven gene prioritization tools that we have analyzed in this study, only Endeavour and Candid have been used for genome-wide prioritization. The input data for Endeavour was training genes, whereas Candid required keywords. The gene prioritization tools that we have analyzed to prioritize candidate genes within a region of interest were SUSPECTS, ToppGene, GeneWanderer, Posmed, GeneDistiller, and again Endeavour. SUSPECTS and Posmed were trained with keywords, the other tools required training genes. We have extensively searched through literature and dedicated databases to identify as many reliable training genes as possible for the disease of interest, as well as a complete set of appropriate keywords in order to be able to derive fair and meaningful comparisons. It is however important to notice that different, and possibly better, results might be obtained by tuning the parameters or by refining the inputs.

Our validation is too small to claim that the differences among the tools are significant. However, a trend can still be observed and GeneDistiller and Endeavour-CS appear as the best tools when looking at the median over the rank ratios (both 11.2%). They are closely followed by SUSPECTS (11.4%). It is interesting to notice that the best results are in general obtained with tools that use many data types in conjunction (up to six and eight respectively for Candid and Endeavour), when compared to SUSPECTS and Posmed (three data types for each). This is in accordance with the conclusion of the recent review by Tiffin et al., who indicate that successful computational applications will be facilitated by improved data integration (Tiffin et al., 2009).

Most of the tools have a high reliability measure ranging from 90% to 100%, meaning that at least 41 of the 42 novel disease genes are prioritized. One notable exception is Posmed, with a reliability of 48.8% and 51.2%, respectively for Posmed-KS and Posmed-DN, which can be explained by the fact that Posmed also acts as a filter on the candidate genes in order to obtain a reduced list of genes in the end. There are therefore cases for which the novel disease gene is not present in the results since it has been removed by the filter. This is different from the other tools for which missing genes basically mean that these genes are not recognized by the tool (it happens most of the time with poorly characterized genes like C20orf54). Another special case is SUSPECTS that went offline during the validation and therefore could only be validated with the first 28 associations. We therefore calculated the reliability only on the first 28 associations.

Two types of tools can be distinguished, the tools that are trained with already known genes and the ones that are trained with descriptive keywords. It is important to notice that gene based methods seem to work better than keyword based methods (average 25.6 for genes based methods (ToppGene, GeneWanderer, GeneDistiller, and Endeavour) and 30.0 for keyword based methods (SUSPECTS, Posmed, GeneDistiller, and Candid)). This can be due to the fact that we used in general more genes than keywords for training (18.8 genes on average for 6.0 keywords). This also indicates that more keywords might be needed to model a disease, even if a small text like the OMIM entry can sometimes be used.

Posmed-KS has been trained with the complete keyword set, whereas Posmed-DN has been trained only with the disease name. There is a difference between the two versions since the median rank ratio is 29,5% when the complete keyword set is used and drops to 37.7% when only the disease

name is inputted. If we only compare the results over the 20 associations for which both methods are able to find back the novel disease gene, the difference becomes even larger (28.8% and 47.7% respectively for Posmed-KS and Posmed-DN). Altogether, these results indicate that Posmed does not rely on the single use of the disease name and that the extra keywords have some impact.

GeneWanderer was also run twice with different network algorithms: random walk (RW) and diffusion kernel (DK). The results indicate that the random walk approach is performing a little bit better than the diffusion kernel albeit non significant (22.0% to 23.3% for median rank ratio - similar differences are observed with the other indicators). The heat map indicates a strong correlation (>0.9, see Figure S1) between the two modes, which was expected since applying diffusion to a kernel can be interpreted as equivalent to applying a random walk on the underlying network. Altogether, this indicates that these two similar network algorithms give similar results and are globally correlated.

Endeavour was used to prioritize both the whole genome (Endeavour-GW) and the defined chromosomal region (Endeavour-CS) allowing us to identify the influence of the size of the gene list to prioritize. The median rank ratio is better for Endeavour-CS (11.2%) than for Endeavour-GW (15.9%). The difference remains when considering the AUC, and the sensitivity at 10%. The same training genes were used, and therefore the observed difference is only caused by extending the small candidate gene set to the whole genome. This confirms previous findings that prioritizing the whole genome is more difficult than prioritizing a rather small positive locus. The heat map indicates that the two Endeavour modes are strongly correlated as expected since the core algorithm is the same in both modes (>0.9, see Figure S1).

An important feature that might influence the results is the date of the last data update. The latest genomic data (still prior to discoveries considered in this study) is likely to give the best results since it will model more accurately what is currently known compared to data that is two year old. In our setup, we have no control over the genomic data used and can not identify if variation in performance among tools can be explained by this.

It is important to notice that the 42 novel disease gene associations do not represent a very homogeneous set (see Table S2). Indeed, among the 42 associations only 23 could be confirmed (in monogenic diseases we expect to find the mutation in at least 2 unrelated patients, and in multifactorial diseases a GWAS is required to be replicated in a separate cohort), 11 associations were weak (not confirmed) associations, and 8 associations were intermediate associations (only 1 study was performed, but there exist additional functional evidence). This influences our validation since some hypothetical associations might in fact be spurious.

## FUNDING

Research Council KUL [ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 *SymBioSys* en KUL PFV/10/016 *SymBioSys* , START 1, several PhD/postdoc & fellow grants]; Flemish Government [FWO: PhD/postdoc grants, projects , G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR), IWT: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3, FOD:Cancer plans, IBBT]; Belgian Federal Science Policy Office [IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011)]; EU-RTD [ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED].

## REFERENCES


[1] C. C. Abnet et al., "A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma," *Nat Genet*, vol. 42, no. 9, pp. 764-767, 2010.

[2] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "SUSPECTS: enabling fast and effective prioritization of positional candidates," *Bioinformatics*, vol. 22, no. 6, pp. 773 -774, Mar. 2006.

[3] S. Aerts et al., "Gene prioritization through genomic data fusion," *Nat Biotech*, vol. 24, no. 5, pp. 537-544, May. 2006.

[4] S. Aerts et al., "Integrating Computational Biology and Forward Genetics in Drosophila," *PLoS Genet*, vol. 5, no. 1, p. e1000351. 2009.

[5] C. B. Arrington, A. Patel, C. A. Bacino, and N. E. Bowles, "Haploinsufficiency of the LIM domain containing preferred translocation partner in lipoma (LPP) gene in patients with tetralogy of Fallot and VACTERL association," *American Journal of Medical Genetics. Part A*, vol. 152, no. 11, pp. 2919-2923, Nov. 2010.

[6] S. Banka, W. G. Newman, R. K. Ozgül, and A. Dursun, "Mutations in the G6PC3 gene cause Dursun syndrome," *American Journal of Medical Genetics. Part A*, vol. 152, no. 10, pp. 2609-2611, Oct. 2010.

[7] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The Genetic Association Database," *Nat Genet*, vol. 36, no. 5, pp. 431-432, May. 2004.

[8] J.-X. Bei et al., "A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci," *Nat Genet*, vol. 42, no. 7, pp. 599-603, Jul. 2010.

[9] F. B. S. Briggs et al., "Evidence for CRHR1 in multiple sclerosis using supervised machine learning and meta-analysis in 12,566 individuals," *Human Molecular Genetics*, vol. 19, no. 21, pp. 4286-4295, Nov. 2010.

[10] C. C. Cardoso et al., "IFNG +874 T>A single nucleotide polymorphism is associated with leprosy among Brazilians," *Human Genetics*, vol. 128, no. 5, pp. 481-490, Nov. 2010.

[11] J. Chen, H. Xu, B. Aronow, and A. Jegga, "Improved human disease candidate gene prioritization using mouse phenotype," *BMC Bioinformatics*, vol. 8, no. 1, p. 392, 2007.

[12] H. S. Cheong et al., "Association of RANBP1 haplotype with smooth pursuit eye movement abnormality," *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, vol. 156, no. 1, pp. 67-71, Jan. 2011.

[13] C. C. Elbers, N. C. Onland-Moret, L. Franke, A. G. Niehoff, Y. T. van der Schouw, and C. Wijmenga, "A strategy to search for common obesity and type 2 diabetes genes," *Trends in Endocrinology and Metabolism: TEM*, vol. 18, no. 1, pp. 19-26, Feb. 2007.

[14] E. Ellinghaus et al., "Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2," *Nat Genet*, vol. 42, no. 11, pp. 991-995, Nov. 2010.

[15] K. Fransen et al., "Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease," *Human Molecular Genetics*, vol. 19, no. 17, pp. 3482-3488, Sep. 2010.

[16] S. Gimelli et al., "Mutations in SOX17 are associated with congenital anomalies of the kidney and the urinary tract," *Human Mutation*, vol. 31, no. 12, pp. 1352-1359, Dec. 2010.

[17] E. K. Green et al., "Variation at the GABAA receptor gene, Rho 1 (GABRR1) associated with susceptibility to bipolar schizoaffective disorder," *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, vol. 153, no. 7, pp. 1347-1349, Oct. 2010.

[18] S. Gretarsdottir et al., "Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm," *Nature Genetics*, vol. 42, no. 8, pp. 692-697, Aug. 2010.

[19] T. B. Haack et al., "Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency," *Nat Genet*, vol. 42, no. 12, pp. 1131-1134, Diciembre. 2010.

[20] S. Haider, B. Ballester, D. Smedley, J. Zhang, P. Rice, and A. Kasprzyk, "BioMart Central Portal--unified access to biological data," *Nucleic Acids Research*, vol. 37, no. Web Server, p. W23-W27, 2009.

[21] G. M. Hirschfield et al., "Variants at IRF5-TNPO3, 17q12-21 and MMEL1 are associated with primary biliary cirrhosis," *Nat Genet*, vol. 42, no. 8, pp. 655-657, 2010.

[22] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey, "Using literature-based discovery to identify disease candidate genes," *International Journal of Medical Informatics*, vol. 74, no. 2-4, pp. 289-298, Mar. 2005.

[23] S. Kantarci et al., "Characterization of the chromosome 1q41q42.12 region, and the candidate gene

DISP1, in patients with CDH," *American Journal of Medical Genetics. Part A*, vol. 152, no. 10, pp. 2493-2504, Oct. 2010.
[24] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *American Journal of Human Genetics*, vol. 82, no. 4, pp. 949-958, Apr. 2008.
[25] A. T. Kraja, H. L. McLeod, M. A. Province, and J. E. Hutz, "CANDID: a flexible method for prioritizing candidate genes for complex human traits.," *Genetic Epidemiology*, vol. 32, no. 8, pp. 779-790.
[26] A. Letra et al., "Follow-up association studies of chromosome region 9q and nonsyndromic cleft lip/palate," *American Journal of Medical Genetics. Part A*, vol. 152, no. 7, pp. 1701-1710, Jul. 2010.
[27] X. Liu et al., "Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis," *Nature Genetics*, vol. 42, no. 8, pp. 658-660, Aug. 2010.
[28] J.R. Lupski et al., "Whole-genome sequencing in a patient with Charcot-Marie-Tooth Neuropathy". *The New England Journal of Medicine*, vol. 362, no. 13, pp. 1181–1191, April 2010.
[29] V. B. Mahajan et al., "Collagen XVIII mutation in Knobloch syndrome with acute lymphoblastic leukemia," *American Journal of Medical Genetics. Part A*, vol. 152, no. 11, pp. 2875-2879, Nov. 2010.
[30] D. M. McDonald-McGinn, H. Feret, H.-D. Nah, S. P. Bartlett, L. A. Whitaker, and E. H. Zackai, "Metopic craniosynostosis due to mutations in GLI3: A novel association," *American Journal of Medical Genetics. Part A*, vol. 152, no. 7, pp. 1654-1660, Jul. 2010.
[31] D. P. B. McGovern et al., "Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease," *Human Molecular Genetics*, vol. 19, no. 17, pp. 3468-3476, Sep. 2010.
[32] V. A. McKusick, *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*, 12th ed. The Johns Hopkins University Press, 1998.
[33] H. C. Mefford et al., "Copy number variation analysis in single-suture craniosynostosis: multiple rare variants including RUNX2 duplication in two cousins with metopic craniosynostosis," *American Journal of Medical Genetics. Part A*, vol. 152, no. 9, pp. 2203-2210, Sep. 2010.
[34] D. Miki et al., "Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations," *Nat Genet*, vol. 42, no. 10, pp. 893-896, Oct. 2010.
[35] N. Miyake et al., "Loss-of-function mutations of CHST14 in a new type of Ehlers-Danlos syndrome," *Human Mutation*, vol. 31, no. 8, pp. 966-974, Aug. 2010.
[36] A. K. Nicholas et al., "WDR62 is associated with the spindle pole and is mutated in human microcephaly," *Nat Genet*, vol. 42, no. 11, pp. 1010-1014, Nov. 2010.
[37] G. Nikoloski et al., "Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes," *Nat Genet*, vol. 42, no. 8, pp. 665-667, 2010.
[38] D. Nitsch, J. Goncalves, F. Ojeda, B. de Moor, and Y. Moreau, "Candidate gene prioritization by network analysis of differential expression using machine learning approaches," *BMC Bioinformatics*, vol. 11, no. 1, p. 460, 2010.
[39] E. A. Otto et al., "Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy," *Nat Genet*, vol. 42, no. 10, pp. 840-850, Oct. 2010.
[40] K. Qidwai et al., "Deletions of Xp provide evidence for the role of holocytochrome C-type synthase (HCCS) in congenital diaphragmatic hernia," *American Journal of Medical Genetics. Part A*, vol. 152, no. 6, pp. 1588-1590, Jun. 2010.
[41] F. C. Radio et al., "TBX2 gene duplication associated with complex heart defect and skeletal malformations," *American Journal of Medical Genetics. Part A*, vol. 152, no. 8, pp. 2061-2066, Aug. 2010.
[42] E. F. Remmers et al., "Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behcet's disease," *Nat Genet*, vol. 42, no. 8, pp. 698-702, 2010.
[43] E. F. Remmers et al., "Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R/IL12RB2 regions associated with Behçet's disease," *Nature genetics*, vol. 42, no. 8, pp. 698-702, Aug. 2010.
[44] M. G. Sampson et al., "Evidence for a recurrent microdeletion at chromosome 16p11.2 associated with congenital anomalies of the kidney and urinary tract (CAKUT) and Hirschsprung disease," *American Journal of Medical Genetics. Part A*, vol. 152, no. 10, pp. 2618-2622, Oct. 2010.
[45] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nat Meth*, vol. 5, no. 1, pp. 16-18. 2008.
[46] D. Seelow, J. M. Schwarz, and M. Schuelke, "GeneDistiller—Distilling Candidate Genes from Linkage Intervals," *PLoS ONE*, vol. 3, no. 12, p. e3874, December. 2008.
[47] V. L. Sheen, A. R. Torres, X. Du, B. Barry, C. A. Walsh, and V. E. Kimonis, "Mutation in PQBP1 is associated with periventricular heterotopia," *American Journal of Medical Genetics. Part A*, vol. 152, no. 11, pp. 2888-2890, Nov. 2010.

[48] E. K. Shin et al., "Association between colony-stimulating factor 1 receptor gene polymorphisms and asthma risk," *Human Genetics*, vol. 128, no. 3, pp. 293-302, Sep. 2010.
[49] A.-C. Tabet et al., "Molecular characterization of a de novo 6q24.2q25.3 duplication interrupting UTRN in a patient with arthrogryposis," *American Journal of Medical Genetics Part A*, vol. 152, no. 7, pp. 1781-1788, 2010.
[50] E. T. Teber, J. Y. Liu, S. Ballouz, D. Fatkin, and M. A. Wouters, "Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies," *BMC Bioinformatics*, vol. 10 Suppl 1, p. S69, 2009.
[51] M. Tesli et al., "Association analysis of PALB2 and BRCA2 in bipolar disorder and schizophrenia in a scandinavian case-control sample," *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, vol. 153, no. 7, pp. 1276-1282, Oct. 2010.
[52] T. A. Thornblad, K. S. Elliott, J. Jowett, and P. M. Visscher, "Prioritization of positional candidate genes using multiple web-based software tools," *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies*, vol. 10, no. 6, pp. 861-870, Dec. 2007.
[53] N. Tiffin et al., "Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes," *Nucleic Acids Research*, vol. 34, no. 10, pp. 3067-3081, 2006.
[54] N. Tiffin, M. A. Andrade-Navarro, and C. Perez-Iratxeta, "Linking genes to diseases: it's all in the data," *Genome Medicine*, vol. 1, no. 8, p. 77, 2009.
[55] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau, "A guide to web tools to prioritize candidate genes," *Briefings in Bioinformatics*, Mar. 2010.
[56] C. Turnbull et al., "Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer," *Nature Genetics*, vol. 42, no. 7, pp. 604-607, 2010.
[57] S. Van Vooren et al., "Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations," *Nucleic Acids Research*, vol. 35, no. 8, pp. 2533-2543, Apr. 2007.
[58] L.-D. Wang et al., "Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54," *Nat Genet*, vol. 42, no. 9, pp. 759-763, 2010.
[59] Q. Xiong, Y. Qiu, and W. Gu, "PGMapper: a web-based tool linking phenotype to genes," *Bioinformatics*, vol. 24, no. 10, p. 1323, May. 2008.
[60] T. Yamauchi et al., "A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B," *Nat Genet*, vol. 42, no. 10, pp. 864-868, Oct. 2010.
[61] Y. Yoshida et al., "PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning," *Nucleic Acids Research*, vol. 37, no. Web Server issue, pp. W147-152, Jul. 2009.
[62] T. W. Yu et al., "Mutations in WDR62, encoding a centrosome-associated protein, cause microcephaly with simplified gyri and abnormal cortical architecture," *Nat Genet*, vol. 42, no. 11, pp. 1015-1020, Nov. 2010.
[63] W. Yu, A. Wulf, T. Liu, M. J. Khoury, and M. Gwinn, "Gene Prospector: An evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases," *BMC Bioinformatics*, vol. 9, pp. 528-528.
[64] M. Zlojutro et al., "Genome-wide association study of theta band event-related oscillations identifies serotonin receptor gene HTR7 influencing risk of alcohol dependence," *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, Nov. 2010.
[65] S. Zucchelli et al., "TRAF6 promotes atypical ubiquitination of mutant DJ-1 and alpha-synuclein and is localized to Lewy bodies in sporadic Parkinson's disease brains," *Human Molecular Genetics*, vol. 19, no. 19, pp. 3759-3770, Oct. 2010.

## TABLES

**Table 1: Ranking positions of the novel disease genes from the validation dataset (in percent) of the genome-wide gene prioritization using CANDID and Endeavour.**

| | | Candid | Endeavour | Average Ranking position |
|---|---|---|---|---|
| **HCCS** | Qidwai et al. (2010) | 25.8 | 2.7 | **14.3** |
| **BRCA2** | Tesli et al. (2010) | 1.3 | 1.2 | **1.3** |
| **TNFRSF19** | Bei et al. (2010) | 5.7 | 36.1 | **20.9** |
| **MECOM** | Bei et al. (2010) | 11.6 | 15.9 | **13.7** |
| **ATF7IP** | Turnbull et al. (2010) | 1.4 | 49.9 | **25.6** |
| **DMRT1** | Turnbull et al. (2010) | 41.8 | 79.0 | **60.4** |
| **FUT2** | McGovern et al. (2010) | 47.7 | 42.4 | **45.0** |
| **CSF1R** | Shin et al. (2010) | 2.9 | 1.7 | **2.3** |
| **GLI3** | McDonald-McGinn et al. (2010) | 7.9 | 2.2 | **5.1** |
| **STOM** | Letra et al. (2010) | 26.5 | 6.5 | **16.5** |
| **UTRN** | Tabet et al. (2010) | 18.5 | 0.4 | **9.5** |
| **GABRR1** | Green et al. (2010) | 17.4 | 15.9 | **16.7** |
| **UBE2L3** | Fransen et al. (2010) | 49.1 | 71.2 | **60.2** |
| **BCL3** | Fransen et al. (2010) | 60.2 | 4.7 | **32.5** |
| **EZH2** | Nikoloski et al. (2010) | 4.6 | 34.4 | **19.5** |
| **TRAF6** | Zucchelli et al. (2010) | 0.1 | 16.5 | **8.3** |
| **IL10** | Remmers et al. (2010), Mizuki et al. (2010) | 13.2 | 26.2 | **19.7** |
| **DAB2IP** | Gretarsdottir et al. (2010) | 14.2 | 26.4 | **20.3** |
| **SPIB** | Liu et al. (2010) | 5.5 | 10.3 | **7.9** |
| **MMEL1** | Hirschfield et al. (2010) | 25.2 | 39.4 | **32.2** |
| **TBX2** | Radio et al. (2010) | 31.3 | 1.8 | **16.6** |
| **RUNX2** | Mefford et al. (2010) | 1.8 | 3.1 | **2.4** |
| **CRHR1** | Briggs et al. (2010) | 15.0 | 13.3 | **14.2** |
| **IFNG** | Cardoso et al. (2010) | 47.2 | 1.2 | **24.2** |
| **SH2B1** | Sampson et al. (2010) | 3.5 | 10.8 | **7.2** |
| **DISP1** | Kantarci et al. (2010) | 2.2 | 32.8 | **17.6** |
| **G6PC3** | Banka et al. (2010) | 47.7 | 18.7 | **33.2** |
| **PQBP1** | Sheen et al. (2010) | 22.5 | 16.7 | **19.6** |
| **CD320** | Quadros et al. (2010) | 80.5 | 46.9 | **63.7** |
| **CHST14** | Miyake et al. (2010) | 61.6 | 7.5 | **34.6** |
| **PLCE1** | Wang et al. (2010) ,Abnet et al. (2010) | 10.4 | 13.0 | **11.7** |
| **C20orf54** | Wang et al. (2010) | 67.5 | 95.5 | **81.5** |
| **SDCCAG8** | Otto et al. (2010) | 78.1 | 5.2 | **41.7** |
| **TP63** | Miki et al. (2010) | 7.0 | 1.3 | **4.2** |
| **UBE2E2** | Yamauchi et al. (2010) | 51.1 | 17.3 | **34.3** |
| **LPP** | Arrington et al (2010) | 74.5 | 6.9 | **40.7** |
| **RANBP1** | Cheong et al. (2010) | 7.3 | 46.0 | **26.7** |
| **HTR7** | Zlojutro et al. (2010) | 17.7 | 0.5 | **9.1** |
| **SOX17** | Gimelli et al. (2010) | 74.6 | 15.1 | **44.9** |
| **ACAD9** | Haack et al. (2010) | 33.4 | 1.5 | **17.5** |
| **TRAF3IP2** | Ellinghaus et al. (2010), Hüffmeier et al. (2010) | 2.0 | 39.0 | **20.5** |
| **WDR62** | Yu et al. (2010), Nicholas et al. (2010) | 29.9 | 23.3 | **26.7** |
| | Rank Ratio (Median) | **17.7** | **15.9** | |
| | Rank Ratio (Mean) | **26.9** | **21.8** | |
| | **Sensitivity 10%** | **15** | **16** | |
| | **Sensitivity 30%** | **28** | **30** | |

**Table 2: Ranking positions of the novel disease genes from the validation dataset of the candidate set based prioritization using SUSPECTS, ToppGene, GeneWanderer (Random Walk and Diffusion Kernel), Posmed (Disease name and Keywords), GeneDistiller, Endeavour.**

| | SUSPECTS | Toppgene | Gene Wanderer (RW) | Gene Wanderer (DK) | Posmed (disease name) | Posmed (keywords) | Gene Distiller | Endeavour | Average Ranking position |
|---|---|---|---|---|---|---|---|---|---|
| **HCCS** | 23.1 | 3.4 | 46.8 | 37.5 | 15.4 | N.A. | 7.7 | 8.9 | **20.4** |
| **BRCA2** | 63.6 | 2.1 | 9.7 | 8.3 | 30.0 | 37.5 | 1.3 | 2.9 | **19.4** |
| **TNFRSF19** | 14.1 | 13.0 | 31.5 | 41.8 | 7.7 | 28.1 | 31.3 | 22.7 | **23.8** |
| **MECOM** | N.A. | N.A. | 45.0 | 11.1 | N.A. | 75.0 | 24.3 | 8.1 | **32.7** |
| **ATF7IP** | 69.9 | 76.8 | 76.5 | 39.5 | N.A. | N.A. | 11.4 | 41.7 | **46.1** |
| **DMRT1** | 0.9 | 71.7 | 28.6 | N.A. | 21.4 | 21.4 | 15.8 | 97.9 | **36.8** |
| **FUT2** | 10.5 | 89.7 | 50.0 | 92.1 | N.A. | N.A. | 26.2 | 27.4 | **49.3** |
| **CSF1R** | 0.9 | 1.6 | 6.3 | 3.8 | 45.4 | 15.5 | 1.1 | 5.0 | **10.0** |
| **GLI3** | 22.1 | 1.1 | 2.0 | 2.4 | 50.0 | 33.3 | 0.8 | 3.8 | **14.5** |
| **STOM** | 17.6 | 67.4 | 6.8 | 4.7 | N.A. | N.A. | 35.6 | 12.7 | **24.2** |
| **UTRN** | 2.5 | 2.9 | 12.5 | 10.5 | 20.0 | N.A. | 5.0 | 2.0 | **7.9** |
| **GABRR1** | 11.4 | 5.2 | 56.0 | 54.5 | 85.7 | 100.0 | 4.5 | 25.7 | **42.9** |
| **UBE2L3** | 79.7 | 87.3 | 16.8 | 16.5 | N.A. | N.A. | 9.1 | 29.0 | **39.7** |
| **BCL3** | 0.8 | 2.2 | 3.6 | 11.2 | N.A. | N.A. | 5.8 | 6.5 | **32.7** |
| **EZH2** | 37.4 | 77.0 | 18.4 | 36.6 | 60.8 | 21.1 | 8.0 | 16.0 | **34.4** |
| **TRAF6** | 4.5 | 40.8 | 6.5 | 11.6 | 4.8 | 23.5 | 0.8 | 27.5 | **5.0** |
| **IL10** | 1.8 | 1.3 | 0.8 | 27.5 | 2.7 | 2.5 | 8.4 | 28.5 | **9.2** |
| **DAB2IP** | 49.3 | 91.2 | 22.0 | 30.2 | N.A. | N.A. | 37.7 | 20.3 | **41.8** |
| **SPIB** | 71.9 | 16.8 | 24.7 | 12.6 | 94.0 | 82.3 | 6.4 | 7.8 | **39.5** |
| **MMEL1** | 4.3 | 58.7 | N.A. | N.A. | N.A. | N.A. | 51.7 | 22.7 | **44.4** |
| **TBX2** | N.A. | 1.1 | N.A. | N.A. | N.A. | N.A. | 11.3 | 0.5 | **4.3** |
| **RUNX2** | 2.3 | 1.6 | 1.1 | 2.9 | 9.5 | 4.2 | 1.3 | 2.0 | **3.1** |
| **CRHR1** | 10.5 | 3.5 | 23.6 | 24.1 | 21.8 | 21.3 | 17.6 | 10.2 | **16.6** |
| **IFNG** | 1.5 | 1.6 | 2.0 | 7.9 | 5.9 | 10.0 | 0.9 | 2.9 | **4.1** |
| **SH2B1** | N.A. | 80.0 | 53.1 | 20.5 | N.A. | N.A. | 7.5 | 11.8 | **34.6** |
| **DISP1** | 67.7 | 10.9 | 11.5 | N.A. | N.A. | N.A. | 8.1 | 22.2 | **24.1** |
| **G6PC3** | 22.8 | 23.8 | 40.4 | N.A. | N.A. | N.A. | 12.0 | 19.1 | **23.6** |
| **PQBP1** | N.A. | 6.5 | 15.2 | 22.9 | N.A. | N.A. | 1.1 | 8.5 | **10.9** |
| **CD320** | N.A. | 51.6 | 74.0 | 85.9 | N.A. | N.A. | 24.0 | 23.7 | **51.8** |
| **CHST14** | N.A. | 30.1 | 20.0 | 97.2 | 100.0 | 100.0 | 25.0 | 7.6 | **54.3** |
| **PLCE1** | N.A. | 3.2 | 22.2 | 35.3 | 60.0 | 29.5 | 27.1 | 11.1 | **26.9** |
| **C20orf54** | N.A. | 36.8 | 98.9 | 95.9 | N.A. | N.A. | N.A. | 97.2 | **82.2** |
| **SDCCAG8** | N.A. | 1.6 | 63.6 | 93.1 | N.A. | N.A. | 40.0 | 1.3 | **39.9** |
| **TP63** | N.A. | 3.2 | 14.3 | 15.1 | N.A. | N.A. | 0.9 | 1.8 | **6.8** |
| **UBE2E2** | N.A. | 94.9 | 51.8 | 52.2 | N.A. | N.A. | 41.5 | 27.4 | **53.8** |
| **LPP** | N.A. | 85.4 | 61.4 | 78.2 | 87.5 | 40.0 | 11.1 | 11.2 | **53.6** |
| **RANBP1** | N.A. | 93.4 | 28.2 | 23.7 | N.A. | N.A. | 18.3 | 57.9 | **44.3** |
| **HTR7** | N.A. | 82.5 | 1.7 | 6.4 | N.A. | N.A. | 0.8 | 0.9 | **7.1** |
| **SOX17** | N.A. | 31.2 | 2.6 | 20.6 | 60.0 | 48.0 | 5.0 | 10.2 | **25.4** |
| **ACAD9** | N.A. | 1.4 | 10.9 | 2.3 | 68.0 | 95.1 | 31.5 | 1.6 | **30.1** |
| **TRAF3IP2** | N.A. | 8.5 | 25.0 | 6.4 | 90.9 | 52.6 | 14.3 | 19.3 | **31.0** |
| **WDR62** | N.A. | 91.5 | 95.7 | 86.7 | N.A. | N.A. | 37.4 | 7.1 | **63.7** |

*Continued from prev. page*

| | SUSPECTS | Toppgene | Gene Wanderer (RW) | Gene Wanderer (DK) | Posmed (disease name) | Posmed (keywords) | Gene Distiller | Endeavour | Average Ranking position |
|---|---|---|---|---|---|---|---|---|---|
| **Rank Ratio (Mean)** | **23.7** | **36.9** | **29.1** | **33.2** | **42.0** | **40.3** | **15.3** | **18.3** | |
| **Reliability (%)** | **89,9 (25/27)** | **95,3 (41/42)** | **95,3 (40/42)** | **88,4 (38/42)** | **51,2 (21/42)** | **48,8 (20/42)** | **97,6 (41/42)** | **100 (42/42)** | |
| **Sensitivity 10%** | **10** | **18** | **11** | **9** | **5** | **3** | **20** | **18** | |
| **Sensitivity 30%** | **18** | **22** | **27** | **22** | **10** | **11** | **34** | **39** | |

## SUPPLEMENTARY TABLES AND FIGURES

**Figure S1: Reliability measures of genome-wide and candidate gene set based prioritization tools.**

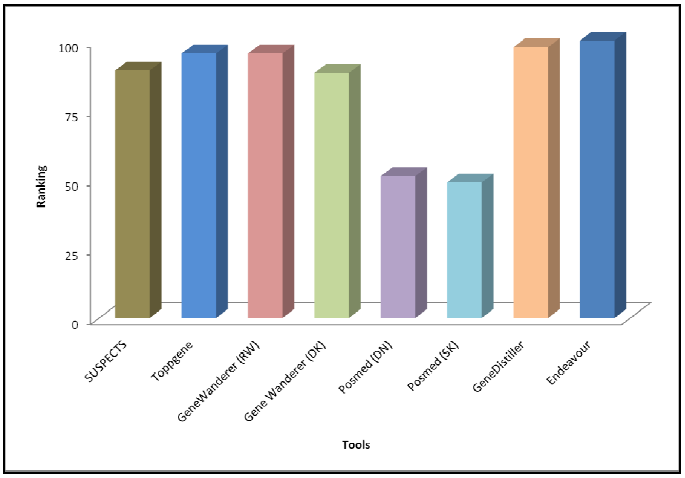



**Figure S2 : Heat Map of the gene prioritization tools used in this study to show their correlation.**

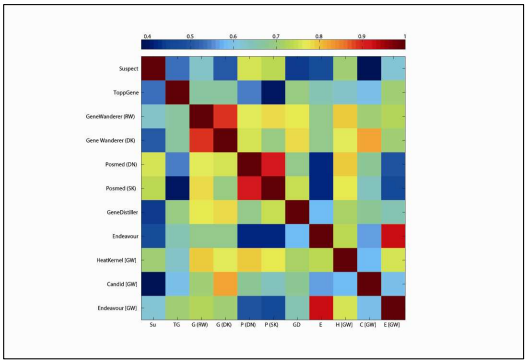


**Table S1 : Overview of the tools for gene prioritization.**

| | GP Tool | Publication | Website |
|---|---|---|---|
| 1 | SUSPECTS | Adie et al. (2006) | http://www.genetics.med.ed.ac.uk/suspects/ |
| 2 | ToppGene | Chen et al. (2007) | http://toppgene.cchmc.org/ |
| 3 | GeneDistiller | Seelow et al. (2008) | http://www.genedistiller.org/ |
| 4 | GeneWanderer | Köhler et al. (2008) | http://compbio.charite.de/genewanderer/GeneWanderer |
| 5 | Posmed | Yoshida et al. (2009) | http://omicspace.riken.jp/ |
| 6 | CANDID | Hutz et al. (2008) | https://dsgweb.wustl.edu/hutz/candid.html |
| 7 | Endevour | Aerts et al. (2006) | http://www.esat.kuleuven.be/endeavour |

**Table S2: Experiments used in this study.**

| Novel Gene | Disease | GAD Tag | Inheritance | Type of study | Finding | Status | Confirmed associations: proof |
|---|---|---|---|---|---|---|---|
| ACAD9 | Mitochondrial complex I deficiency | MITOCHONDRIAL | monogenic | patient report | mutation | confirmed | 4 unrelated patients. |
| CHST14 | Ehlers-Danlos syndrome (EDS) | OTHER | monogenic | patient report | mutation | confirmed | 6 unrelated patients. |
| EZH2 | Myelodysplastic syndromes (MDS) | CANCER | monogenic | mutation screen | deletion, mutation | confirmed | 22 unrelated patients. |
| G6PC3 | Dursun syndrome | HEMATOLOGICAL | monogenic | case report | mutation | confirmed | 2 unrelated patients. |
| GLI3 | Metopic craniosynostosis | OTHER | monogenic | patient report | mutation | confirmed | 2 unrelated patients. |
| HCCS | Congenital Diaphragmatic Hernia (CDH) | OTHER | monogenic | patient report | deletion | confirmed | 4 unrelated patients. CDH can be a feature of MLS. |
| SDCCAG8 | Retinal-renal ciliopathy | RENAL/VISION | monogenic | patient report | mutation | confirmed | 8 unrelated patients. |
| SOX17 | Congenital anomalies of the kidney and the urinary tract (CAKUT) | DEVELOPMENTAL | monogenic | patient report | mutation | confirmed | 8 unrelated patients. |
| WDR62 | autosomal recessive primary microcephaly and microcephaly with simplified gyri | DEVELOPMENTAL | monogenic | GW linkage analysis | SNP | confirmed | 2 unrelated families. |
| CRHR1 | Multiple sclerosis | IMMUNE | multifactorial | GWAS | SNP | confirmed | Replication in independent cohort and meta-analysis. |
| DAB2IP | Abdominal aortic aneurysm | CARDIOVASCULAR | multifactorial | GWAS | SNP | confirmed | Replication in independent cohort and meta-analysis. |
| DMRT1 | Testicular germ cell tumor (TGCT) | CANCER | multifactorial | GWAS | SNP | confirmed | Kanetsky et al. (2011) Hum Mol Genet. 20(15):3109. |
| FUT2 | Crohn's disease (CD) | IMMUNE | multifactorial | GWAS | SNP | confirmed | Replication in independent cohort and GWAS meta-analysis. |
| GABRR1 | Bipolar schizoaffective disorder | PSYCH | multifactorial | GWAS | SNP | confirmed | Wang et al. (2010) Schizophr Res. 124(1-3):192. |
| IFNG | Leprosy | INFECTION | multifactorial | GWAS | SNP | confirmed | Replication in independent cohort. |
| IL10 | Behçet's disease | IMMUNE | multifactorial | GWAS | SNP | confirmed | Replication in independent cohort and meta-analysis. |
| MMEL1 | Primary biliary cirrhosis | METABOLIC | multifactorial | GWAS | SNP | confirmed | Replication in independent cohort and meta-analysis. |
| RANBP1 | smooth pursuit eye movement abnormality | PSYCH | multifactorial | SNP geno-typing | SNP | confirmed | Cheong et al. (2011) Am J Med Genet B Neuropsychiatr Genet. 156B(1):67. |

*Continued from prev. page*

| Novel Gene | Disease | GAD Tag | Inheritance | Type of study | Finding | Status | Confirmed associations: proof |
|---|---|---|---|---|---|---|---|
| PLCE1 | Esophageal squamous cell carcinoma | CANCER | multifactorial | GWAS | SNP | confirmed | Two unrelated studies. |
| SPIB | Primary biliary cirrhosis | METABOLIC | multifactorial | GWAS | SNP | confirmed | Tanaka et al. (2011) Tissue Antigens. 78(1):65. |
| TP63 | Lung adenocarcinoma | CANCER | multifactorial | GWAS | SNP | confirmed | Hu et al. (2011) Nat Genet.43(8):792. |
| TRAF3IP2 | Psoriasis | IMMUNE | multifactorial | GWAS | SNP | confirmed | Two unrelated studies. |
| UBE2E2 | Type 2 diabetes | METABOLIC | multifactorial | GWAS | SNP | confirmed | Replication in two independent cohort. |
| CD320 | Methylmalonic aciduria | METABOLIC | monogenic | Patient report | deletion | intermediate | |
| DISP1 | Congenital Diaphragmatic Hernia (CDH) | OTHER | monogenic | GWAS | deletion, micro-deletion, SNP, mutation | intermediate | |
| PQBP1 | Periventricular heterotopia | UNKNOWN | monogenic | Patient report | deletion | intermediate | |
| UTRN | Arthrogryposis | UNKNOWN | monogenic | Patient report | mutation | intermediate | |
| C20orf54 | Esophageal squamous cell carcinoma | CANCER | multifactorial | GWAS | SNP | intermediate | |
| HTR7 | alcohol dependence | CHEMDEPENDENCY | multifactorial | GWAS | SNP | intermediate | |
| STOM | Nonsyndromic cleft lip/palate | DEVELOPMENTAL | multifactorial | GWAS | SNP | intermediate | |
| UBE2L3 | Crohn's disease | IMMUNE | multifactorial | GWAS | SNP | intermediate | |
| LPP | Tetralogy of Fallot | CARDIOVASCULAR | monogenic | Patient report | deletion | weak | |
| RUNX2 | Single-suture craniosynostosis | OTHER | monogenic | mutation screen | duplication | weak | |
| SH2B1 | Congenital Anomalies of the Kidney and Urinary Tract | DEVELOPMENTAL | monogenic | case report | Micro-deletion | weak | |
| TBX2 | Complex heart defect | CARDIOVASCULAR | monogenic | case report | duplication + model organism data | weak | |
| ATF7IP | Testicular germ cell tumor (TGCT) | CANCER | multifactorial | GWAS | SNP | weak | |
| BCL3 | Crohn's disease | IMMUNE | multifactorial | GWAS | SNP | weak | |
| BRCA2 | Bipolar schizoaffective disorder | PSYCH | multifactorial | GWAS | SNP | weak | |
| CSF1R | Asthma | IMMUNE | multifactorial | GWAS | SNP | weak | |
| MECOM | Nasopharyngeal carcinoma (NPC) | CANCER | multifactorial | GWAS | SNP | weak | |
| TNFRSF19 | Nasopharyngeal carcinoma (NPC) | CANCER | multifactorial | GWAS | SNP | weak | |
| TRAF6 | Parkinson's disease | NEUROLOGICAL | multifactorial | Hypo-thesis | expression and interaction | weak | |

# Chapter 4

# *Study 3*: Gene prioritization by network analysis of expression data

Most of the computational methods for gene prioritization that have been proposed so far, are based on the guilt-by-association concept and are not applicable for unknown diseases or phenotypes. We replace the prior knowledge about the biological process with experimental data on differential gene expression and consider the problem on a network level, assuming that a strong candidate will tend to interact with many differentially expressed neighbors in the network. For doing that, we define a notion of soft neighborhood where each gene is given a contributing weight, which decreases with the distance from the candidate gene on the protein network. To account for multiple paths between genes, we define the distance using the Laplacian exponential diffusion kernel. We score candidates by aggregating the differential expression of neighbors weighted as a function of distance. Through a randomization procedure, we rank candidates by p-values. Finally, we illustrate our approach on four monogenic diseases and successfully prioritize the known disease causing genes.

**Publication**
This work has been published in 2009 as an electronic version in the *PLoS ONE* journal (ISI IF 2009: 4.3).
*Reference:* Nitsch D, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y (2009) Network Analysis of Differential Expression for the Identification of Disease-Causing Genes. *PLoS ONE 4*(5): e5526. PMID: 19436755.

**Personal contribution**

The PhD candidate has designed and implemented the methodology in Matlab, set up and applied the case studies for evaluating the methodology, analyzed the results and drafted the paper. The first co-author set up the network and participated in writing the paper. Biological issues were addressed by the second co-author, and and the biological investigation of the suggested novel candidate gene was done by the third co-author.

OPEN ACCESS Freely available online

PLoS one

# Network Analysis of Differential Expression for the Identification of Disease-Causing Genes

**Daniela Nitsch[1]*, Léon-Charles Tranchevent[1], Bernard Thienpont[2], Lieven Thorrez[1], Hilde Van Esch[2], Koenraad Devriendt[2], Yves Moreau[1]**

**1** Department of Electrical Engineering (ESAT-SCD) Katholieke Universiteit Leuven, Leuven, Belgium, **2** Center for Human Genetics, University Hospitals Leuven, Leuven, Belgium

## Abstract

Genetic studies (in particular linkage and association studies) identify chromosomal regions involved in a disease or phenotype of interest, but those regions often contain many candidate genes, only a few of which can be followed-up for biological validation. Recently, computational methods to identify (prioritize) the most promising candidates within a region have been proposed, but they are usually not applicable to cases where little is known about the phenotype (no or few confirmed disease genes, fragmentary understanding of the biological cascades involved). We seek to overcome this limitation by replacing knowledge about the biological process by experimental data on differential gene expression between affected and healthy individuals. Considering the problem from the perspective of a gene/protein network, we assess a candidate gene by considering the level of differential expression in its *neighborhood* under the assumption that strong candidates will tend to be surrounded by differentially expressed neighbors. We define a notion of *soft* neighborhood where each gene is given a contributing weight, which decreases with the distance from the candidate gene on the protein network. To account for multiple paths between genes, we define the distance using the Laplacian exponential diffusion kernel. We score candidates by aggregating the differential expression of neighbors weighted as a function of distance. Through a randomization procedure, we rank candidates by *p*-values. We illustrate our approach on four monogenic diseases and successfully prioritize the known disease causing genes.

**Citation:** Nitsch D, Tranchevent L-C, Thienpont B, Thorrez L, Van Esch H, et al. (2009) Network Analysis of Differential Expression for the Identification of Disease-Causing Genes. PLoS ONE 4(5): e5526. doi:10.1371/journal.pone.0005526

**Editor:** Gustavo Goldman, Universidade de Sao Paulo, Brazil

**Received** December 19, 2008; **Accepted** April 20, 2009; **Published** May 13, 2009

**Copyright:** © 2009 Nitsch et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Research supported by Research Council KUL [GOA AMBioRICS, CoE EF/05/007 SymBioSys (LT), PROMETA, several PhD/postdoc & fellow grants], Flemish Government [Fonds voor Wetenschappelijk Onderzoek: PhD/postdoc grants, projects G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, research communities (ICCoS, ANMMM, MLDM); HVE and KD are Clinical Investigators supported by FWO], IWT [PhD Grants (BT), GBOU-McKnow-E, GBOU-ANA, TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM-IOTA3, O&O-Dsquare], Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007–2011)], EU-RTD [ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain (LCT, DN), FP6-STREP Strokemap]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: Daniela.Nitsch@esat.kuleuven.be

## Introduction

Genetic studies, including cytogenetic, linkage, and association studies, can identify chromosomal regions associated with a disease or phenotype of interest. Similarly, high-throughput 'omics' experiments identify genes or proteins implicated in a biological process of interest. In both cases, biologists are often confronted with long lists of tens or hundreds of candidate genes among which they need to select a limited number of candidates for further validation. This problem has been termed *gene prioritization* [1]. Recently, computational methods for prioritizing candidate genes have been proposed. They usually rank candidates by matching their information across multiple data sources against a profile derived from a set of genes, pathways, or biological processes already known to be involved in the phenotype. A frequent objection to this class of methods is that they cannot be applied to cases where little is known about the molecular basis of the phenotype (no confirmed disease genes, fragmentary understanding of the biological cascades involved). We seek to overcome this limitation by replacing knowledge about the biological process by experimental data on differential gene expression between affected and healthy individuals. Considering the problem from the perspective of a gene/protein network, we assess a candidate gene by considering the level of differential expression in its *neighborhood* under the assumption that strong candidates will tend to be surrounded by differentially expressed neighbors.

A number of methods are currently available for gene prioritization. *Aerts et al. (2006)* developed ENDEAVOUR, a gene prioritization method that ranks candidate genes based on their similarity to genes already involved in the biological process of interest, using multiple data sources (e.g., sequence, expression, literature) [2]. Similarly, *Köhler et al. (2008)* developed GeneWanderer, a method for prioritizing candidate genes by the use of the random walk analysis that defined similarities in protein-protein interactions (PPI) networks [3]. Their global distance measure defines the similarity between genes within the global network and ranks candidate genes on the basis of their similarity to known disease genes. They hypothesized that a global network-similarity measure captures associations between disease proteins better than algorithms based on direct interactions or shortest paths between disease genes. *Franke et al. (2006)* also incorporated the interactions between genes in a network to prioritize candidate genes [4]. They

developed a human gene network that integrates information on genes and their functions. Their method, Prioritizer, ranks candidate genes on the basis of their interactions. They analyze susceptibility loci and investigate whether genes from different loci can be linked to each other. *Lage et al. (2007)* developed a phenome-interactome network that integrates phenotypic literature information from OMIM with a cross-species PPI network [5]. They implemented a Bayesian disease gene predictor that computes for each candidate gene the probability that it is the disease-related gene. High probabilities are assigned to genes that interact with genes that are already associated with phenotypically-related disorders. *Chuang et al. (2007)* developed a network analysis method by applying a protein network-based approach that identifies biomarkers not as individual genes but as subnetworks extracted from protein interaction databases [6]. To find associations between phenotypes and subnetworks, they developed a scoring based on mutual information measure. Although their methodology resembles that of gene prioritization, their method is related to biomarker discovery rather than prioritization.

Most of available prioritization methods have in common that they require knowledge about the disease to identify putative disease genes, for example in the form of a set of genes, pathways, or gene ontology categories known to be implicated in the disease. They then rank candidate genes through "guilt by association" methods across a variety of data sources. But when nothing or only little is known about a disease, these methods will be inapplicable or ineffective. While it is extremely useful to incrementally add disease genes to phenotypes whose molecular basis is reasonably well characterized, there is strong demand from geneticists for methods that could help in the more difficult case of disorders for which the molecular basis is not yet elucidated. Currently, there are not yet truly effective prioritization methods for this case.

We seek to overcome this limitation by replacing knowledge about the biological process by experimental data on differential gene expression between affected and healthy individuals. Our method is a generalization, from a systems biology perspective, of a standard procedure for assessing candidate genes in genetic studies. A standard genetic procedure to analyze candidate genes is to check the expression level of a candidate gene in patient-derived material against wild type (typically in fibroblast or immortalized lymphoblastoid cell lines). Candidates for which a significant difference is observed between the two groups are considered promising. However, for many genetic diseases (such as diseases arising from point mutations in coding regions), there is no guarantee that the expression level of the disease gene itself is affected (although this is possible through feedback effects). Rather, genes "downstream" of the disease gene are those whose expression will be affected. Instead of considering genes in isolation, we consider the differential expression data now at the level of a gene/protein network. If we look at expression patterns mapped on a gene network, we therefore expect that we will observe a disrupted expression module around the disease gene. Other candidate genes, which are not causally related to the phenotype, should not be part of such a disrupted expression module. For this reason, the entire affected neighborhood has to be considered for each candidate gene instead of only taking its own expression level into account.

When considering the expression data at the network level, we need to rely on a gene/protein network. Originally, protein-protein interaction networks from data on putative physical interaction between proteins. More recently, protein networks combining a variety of information sources have been proposed. A link in such a network does not necessarily imply physical interaction between two proteins, but rather some form of association resulting from different types of data (actual interaction, membership in the same pathway, coexpression, literature cooccurrence, etc.) We will call such a network, a protein association network, or protein network for short. Note that in such networks no distinction is made between gene and protein, or multiple isoforms. Furthermore, protein networks are far from complete, and dealing with direct protein-protein interactions may be suboptimal since protein networks are still sparse because of many unknown components and pathways [4,7]. Also, procedures that define neighborhoods in terms of the minimum number of steps from a given gene suffer from the "small world" effect (i.e., the number of neighbors of a given gene grows rapidly with the number of steps taken along the network). To overcome those potential limitations, we chose to use a global distance network that considers both direct and indirect paths in the network [3]. By considering indirect interactions in a protein network, missing links in the network can be compensated. Thus, a global distance network is more densely connected than the sparse protein network from which it is derived. Specifically, our method derives the global distance network from a functional protein association network (STRING [8]) using kernel methods. STRING integrates both protein-protein interactions as well as predicted interactions based on comparative genomics and text mining [8].

Figure 1 shows an overview of our approach. For each gene in the network, its differential expression level is determined by transcriptome-wide microarray experiments of mutant vs. wild-type lines. Then differentially expressed neighborhoods in this network are considered for all candidate genes from a chromosomal region of interest (e.g., identified in a linkage study). Candidate genes with genes in their neighborhood having highly differentially expressed levels are strong candidates. The neighborhood of a candidate gene is defined by direct connections in the global distance network, whereby its size can vary. The smaller the distance from a neighboring gene to the candidate, the closer the neighboring gene is in the network. We have chosen to work with the notion of a *soft neighborhood* by which we mean that the neighborhood of a gene is not a limited set of gene, but rather a weighting function that decreases monotonically as a function of the distance from the gene, but potentially covers the whole network (this idea is reminiscent of the idea of fuzzy membership in fuzzy c-means clustering).

To identify candidates belonging to a significantly disrupted expression module, we have developed a novel randomization method that identifies modules with significantly affected genes. Each candidate receives a score based on the analysis of the differential expression along its neighborhood. The level of differential expression of each gene is weighted by its network distance from the candidate and summed up over all genes. Therefore, the higher the differential expression levels of neighboring genes with small distances are, the higher the score. To determine a candidate gene's significance, a $p$-value based on a randomization procedure is computed. If a candidate belongs to a significant disrupted expression module in the network, its $p$-value is expected to be significant.

Evaluating such a method *in silico* is obviously challenging because if we make predictions on actual diseases where no genes are known, the only way to validate those predictions will be to carry out a full biological validation. Moreover, the kind of expression data needed for our analysis will be currently available only for very few diseases because at this point biologists mostly carry out this type of experiments for diseases for which the cause is known. In a first step, we therefore mimic the situation by taking known disease genes for which expression data for patients versus controls is available and attempt to recover the gene mutated in

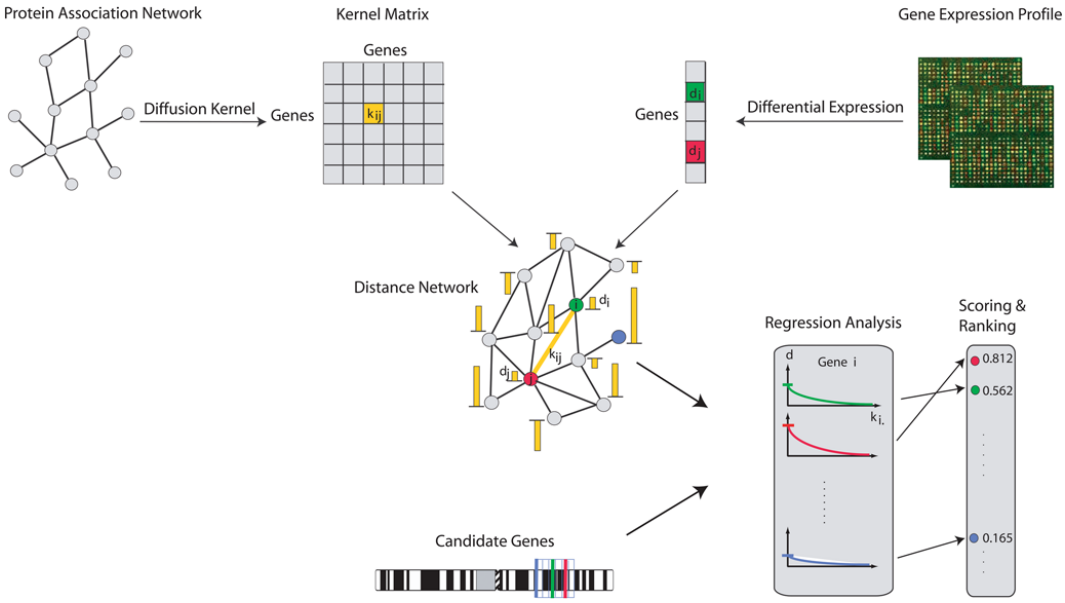



**Figure 1. Overview of the method.** (1) From a protein association network a global distance network is computed based on a kernel method (e.g., the Laplacian exponential diffusion kernel). (2) A disease related microarray experiment is required from which the differential expression levels of all genes in the network are determined. (3) The differential expression levels of the genes are mapped on the global distance network. (4) A set of candidate genes is required (e.g., from a linkage study). (5) For each candidate gene its differentially expressed neighborhood is identified by a regression analysis. (6) Based on the regression analysis each candidate gene receives a score. (7) The candidate genes are ranked by their scores.
doi:10.1371/journal.pone.0005526.g001

those patients. We present four distinct data sets for which we could successfully prioritize the disease-causing gene for each data set. The average rank of the known disease genes was approx. 4 out of approx. 120 candidate genes (see Results). Thus, our method can detect unknown disease causing genes by identifying the most disrupted expression modules in the global distance network. In a second step, using one of the only available human expression data sets for a disorder of unclear etiology, we have applied our method to the polygenic disorder Stein-Levental syndrome for which no disease gene is known. We could confirm the influence of two candidate genes: fibrillin 3 (FBN3, for which a susceptibility allele has been identified) and follistatin (FST, for which association with a Stein-Levental syndrome related metabolic phenotype has been shown). Finally, we suggest a new candidate gene (DEAD box 4) potentially involved in this disease.

## Results

Our distance network was derived from the STRING database [8], from which we used all data types provided (genomic context, high-throughput experiments, coexpression data, and previous knowledge). The resulting network detects all direct and indirect connections of genes and represents a notion of global distance measure (see Materials and Methods).

We illustrate our method by its application to the analysis of constitutional genetic disorders caused by a single gene mutation. Following the current practice of assessing candidate gene expression in EBV cell lines or fibroblast cultures in mutant against wild type, we consider expression data from such biological material (or other accessible tissue biopsies). There is however only a limited number of such data sets publicly available through the ArrayExpress [9] and Gene Expression Omnibus [10] repositories. We present here the results of the method on four case studies.

We distributed all signals randomly over the network to estimate the significance of the candidates. In an adequate data set, at least one gene should be found with a significant p-value (i.e., $p<0.05$). We then assessed how high the actual disease gene ranked and whether its score was significant.

### Case studies

We have evaluated our approach on four data sets: (1) fragile X syndrome (FXS) [11] caused by mutation of FMR1 (fragile X mental retardation 1), (2) Marfan syndrome (MFS) [12] caused by mutation of FBN1 (fibrillin 1), (3) cystic fibrosis (CF) [13] caused by mutation of CFTR, and (4) Becker muscular dystrophy (BMD) [14] caused by mutation of DMD.

For each data set, we have determined a set of candidate genes by taking the genes within a set of chromosomal bands centered on the disease-causing or disease-related gene to gather approx. 120 genes. Genes that were absent in our distance network were not further considered. For finding differentially expressed genes in the network we computed the fold-change for each gene in the genome (see Materials and Methods). Beside the scores of the candidates and the fold-change derived from the microarray experiments, we also present known links to similar diseases with related phenotypes.

Each data set and the high-ranking candidates (i.e., those that have a significant p-value) that are phenotypically linked to related diseases [15–21] are characterized in the Supplementary Materials S1. This demonstrates the significance of the method and shows

that not only the actual disease-causing gene can be identified, but also related genes that may also be involved in this disease.

### Neighborhood

Determining an adequate size for the neighborhood that influences the score of a candidate is a challenge and influences the ranking. For three data sets (fragile X syndrome, Marfan syndrome and Cystic fibrosis) we have determined small neighborhoods of 150 or 20 neighboring genes, because for these data sets we obtained the best signal for small neighborhoods (data not shown) after applying the Fisher omnibus statistics (see Materials and Methods for more details). However, for one data set (Becker muscular dystrophy) we have determined a larger neighborhood of 2000 neighboring genes due to high signal for a large neighborhood (data not shown).

To illustrate the difference between disrupted expression neighborhoods of significant candidate genes and not significant genes, we have added graphs containing the queried neighborhood of the candidates (Figures S1, S2, S3). These graphs show the differential expression levels of the neighbors and their distances to the candidate gene. We can observe that the closest neighborhood of a significant candidate is highly differentially expressed and belongs to a more disrupted expression module than the neighborhood of a nonsignificant candidate.

### Significance

To determine a candidate gene's significance, the differential expression levels are distributed randomly over the network. The candidate's score is compared with the distribution of the randomly generated scores that leads to a $p$-value. If a candidate's score is larger than 95% of all randomized scores, this candidate gene can be considered as belonging to a significant disrupted expression module in the network.

Figure 2 shows for all data sets the distribution of $p$-values after 10,000 randomizations, and the $p$-values of the disease genes whereby all were assigned significant $p$-values. These plots demonstrate that a clear distinction could be made between significant genes with low $p$-values and all other genes, and that the disease genes could be identified by their significant $p$-values and their high ranking.

### Ranking

Tables S1, S2, S3, S4 show the results of the ranking of the evaluation data sets. In all four benchmark data sets, the disease causing genes were ranked in the top 10 (FXS: 1$^{st}$ position, MFS: 5$^{th}$ position, CF: 7$^{th}$ position, BMD: 2$^{nd}$ position) out of lists containing approximately 120 candidate genes (see Supplementary Material), and all were assigned significant $p$-values. For all data sets, we could identify several disrupted expression modules of different sizes around a candidate gene. For two data sets (FXS and CF), some of the top ranked genes were already known to be directly associated with a related disease or phenotype [15–21] that emphasizes the significance of this result. We could not only successfully identify the disease genes, but also genes that correlate with the corresponding phenotype.

One of our aims was to develop a method that is independent of the differential expression levels of a candidate gene itself. Therefore, for evaluation purposes, we did *not* take the differential expression levels of the candidates into account but, preferably, the levels of their neighboring genes. Although in practice, we should obviously take into account the differential expression level of the candidate itself (because the disease gene can be disrupted through feedback effects). Our rankings in Tables S1, S2, S3, S4 demonstrate clearly that the ranking orders do not depend on the up- or down-regulation of the candidate genes themselves, but rather on the effects of being positioned in their neighborhoods. Thus, genes that are not differentially expressed can rank higher than highly differentially expressed genes as long as their neighborhood is differentially expressed.

### Application to Stein-Levental syndrome

We have applied our approach to the Stein-Levental syndrome [22], which is characterized by obesity, hyperandrogenism, and chronic anovulation in women. Stein-Levental syndrome is an endocrine disorder that affects approximately 5% of women and is a leading cause of infertility. This syndrome is believed to be oligogenic (i.e., caused by the effect of a limited number of genes) rather than monogenic [23]. Follistatin (FST) was originally proposed as a candidate for Stein-Levental syndrome by linkage and association studies [24], but is now thought to be rather associated with key androgenic phenotypes of Stein-Levental syndrome but not with the disease itself. These results suggest the existence of another disease-causing gene for Stein-Levental syndrome in the vicinity of FST (rather than FST itself). Recently, a Stein-Levental syndrome susceptibility locus was identified at 19p13.2 [25] and further association studies have suggested fibrillin 3 (FBN3) as the Stein-Levental syndrome susceptibility locus [26]. However, beside FST and FBN3 other candidate genes have been studied, often with inconsistent results [27].

We have determined two sets of candidate genes located on the chromosomal region of FST (chr5q11.2) and FBN3 (chr19p13.2). Both genes (FST, FBN3) were ranked high by our method (Tables S5, S6) and we could confirm their important role in this disease. We further detected the DEAD box 4 gene that is located on chr5q11.2 and that was ranked on the top position with a significant $p$-value (Table S5). We suggest that DEAD box 4 is a new candidate gene for Stein-Levental syndrome because of association with stem cell recruitment to the ovaries, interaction with the microRNA processing machinery, and impact on apoptosis [28–35].

## Discussion

As mentioned in the introduction, there are several known methods to prioritize candidate genes. These approaches can be split into methods that need known disease-gene associations about the disease [2,3,5] and methods without this precondition [4]. If little knowledge is available for a specific disease, the methods that require known disease genes will be ineffective. *Franke et al. (2006)* presented a method to prioritize candidate genes without a training set [4], but without using expression data to evaluate the candidate genes. Among a set of disease loci, they will look for pairs, triples, etc. of genes at different loci for which interaction has been described. Such sets of interacting genes are considered more likely candidates for causing the disease at the different loci. Because of the combinatorial nature of the test, the method does not have high statistical power. Along other lines, *Chuang et al. (2007)* introduced a network analysis method using expression data [6]. However, they focused on finding active subnetworks and biomarkers in cancer and developing a methodology that is not directly related to our problem. Our strategy differs as we focused on searching for disease genes for which genetic mutation causes constitutional disorders.

The strength and uniqueness of our approach is that we substitute disease-specific experimental data (in our case expression data) for the prior knowledge of the molecular basis of the disorder. ENDEAVOUR [2] could also incorporate disease-specific expression data but did not use any network analysis

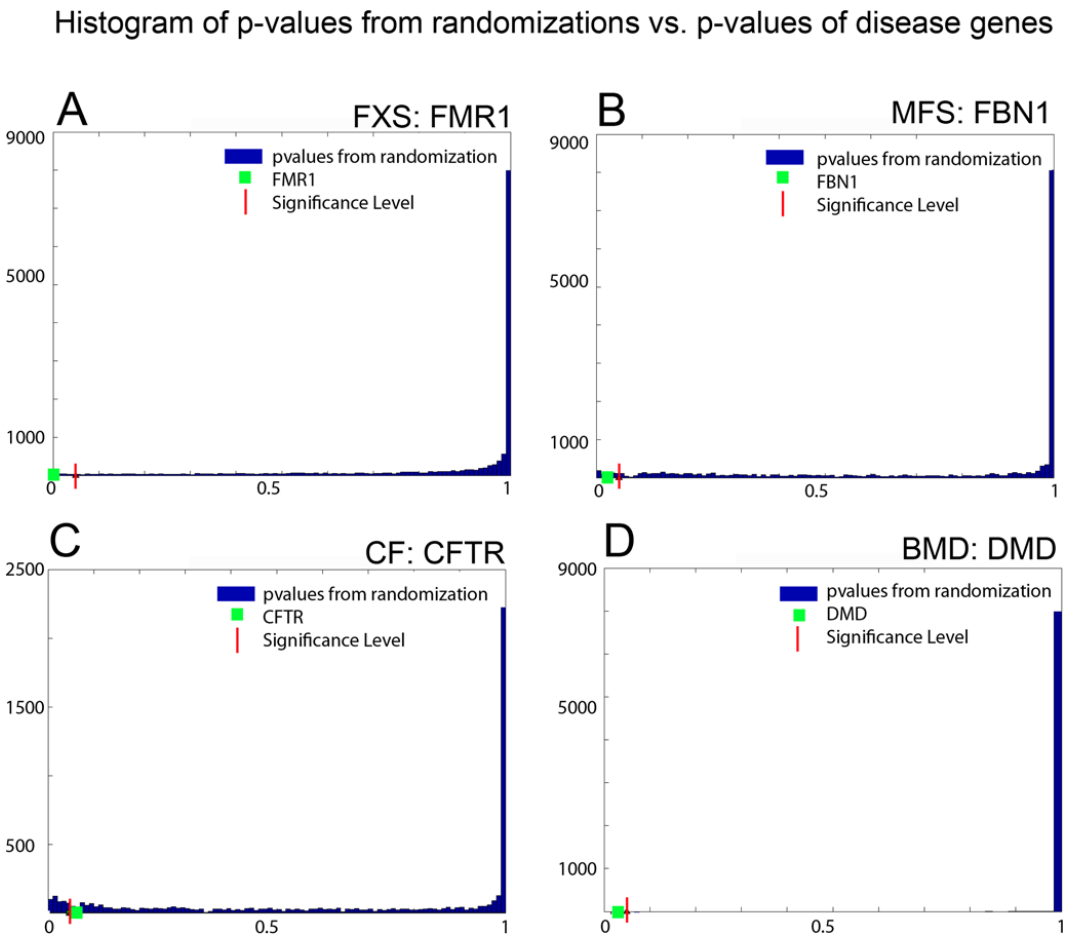



**Figure 2. Histogram of *p*-values from randomizations vs. *p*-values of the disease genes FMR1, FBN1, CFTR, and DMD.** This figure displays the histogram of *p*-values obtained after 10,000 randomizations, and the *p*-values of the disease-causing genes from all four benchmark data sets. It can be observed that in all data sets only a small set of genes have significant *p*-values (i.e., they belong to disrupted expression modules) and all disease causing genes are among them.
doi:10.1371/journal.pone.0005526.g002

concepts and would still mostly rely on the knowledge from the set of known genes for the disease. Among all other methods, only *Chuang et al. (2007)* leverages experimental data [6], however with an entirely different scope and method. Essentially, although existing methods could be applied on the known benchmark cases, none of them would be directly applicable to the actual situation where we want to use our method, which is when the molecular basis of the disease is unknown (with the case of Stein-Levental syndrome as an illustration).

A first question is in which setups our method is applicable. We have chosen here to apply our method in a setup where a locus is known for the disease. Although this is not a strict requirement and genes can be prioritized on a genomewide basis, it has the advantage of limiting the number of candidates tested and therefore limiting the number of false positives. It also guarantees that at least one gene must be associated with the disease. However, among genetic studies, while the method is relevant to linkage and association studies, it may not be applicable to loci detected by cytogenetic studies of patients with genomic deletions or duplications. Indeed, in this case, multiple genes are affected by a copy number change, so that the expression data can be expected to be the superposition of the downstream effects of all the affected genes (although it could be that most of the phenotype is explained by a single critical gene that dominates the downstream cascade of expression dysregulation, in which case the method may still be applicable). Although the concept of our method may seem at first tailored to monogenic disorders, it is more broadly applicable. This is demonstrated by our case study on the polygenic disorder Stein-Levental syndrome, where we could detect FBN3 and FST as related to the disease.

Several factors influence the performance of our prioritization method. First, the quality and coverage of the network around the

actual disease gene will be a strict bottleneck. For example, an isolated gene with no edges in the network can never be effectively prioritized by our method (except by relying solely on its own expression data). We have chosen to use the STRING database (version 7.1) because it is built by taking into account multiple heterogeneous data sources. This results in a large network with a good coverage (human: 16,050 genes, mouse: 16,566 genes). Errors in the network (caused by incorrect gene annotation, unreliable functional annotation, etc.) may both cause false negatives (missing the disease gene) as false positives (genes incorrectly identified as promising because of incorrect association with other genes). Continuing improvements in the quality of protein association networks will contribute to increased effectiveness of the proposed method. Moreover, protein association networks are naïve in terms of alternative transcripts of genes and protein isoforms. At this moment, no distinction is possible among them. It is also unclear what the quality and coverage of nonprotein-coding genes is in current protein association networks – although our candidate gene DEAD box 4 for Stein-Levental syndrome suggests a possible role for the nonprotein-coding genes PIWIL2 (MILI) and DICER1. Several alternatives exist to the STRING network including BioGRID, IntAct, IntNetDb, HPRD, and Dip, and we will study the impact on performance of the choice of network in follow-up work.

Second, the quality and the relevance of the expression data will greatly influence the results of the method. Poorly collected samples (patients not actually sharing the same molecular phenotype) or disease heterogeneity (multiple genes or pathways leading to similar phenotypes) will obscure the expression pattern and will make picking up a meaningful signal more difficult. Similarly, lack of access to the most relevant cell types (biopsies cannot be performed arbitrarily on patients) can be a limiting factor for the method. If the relevant molecular cascades are simply not active in the cell type assessed, differences in expression may be meaningless. The idea is that the pattern of differential expression is as concentrated as possible on the network so that the affected subnetworks can be effectively identified. Some mutations may lead to extensive downstream cascades that may be reflected in network patterns that are too broad to effectively pinpoint the disease-causing gene. A promising experimental direction for focusing the expression patterns towards the disease-causing gene would be to use more sophisticated factorial disease for the microarray experiment. If we can identify a stimulus that is incorrectly processed in affected individuals (e.g., a metabolite or a protein), we could attempt to perform an expression profiling experiment where patient and control material receive a treatment that triggers the affected cascades. The differential expression response will tend be more tightly focused towards those genes that are essential for the difference in response to the treatment between affected vs. non-affected individuals.

Third, an important parameter of the method is the neighborhood size or, the scale parameter of our weighting function that defines it. To study the neighborhood of the candidate genes in order to identify disrupted expression modules in the distance network, we had to bound the neighborhood to a limited size because the network was very densely connected, and thus the neighborhood of a candidate gene consisted of almost all genes in the network. This limitation was done by only considering the neighboring genes with the smallest distances to the candidate (any gene further than a certain threshold was not considered). We wanted the size of the neighborhood to be dependent on the disruption we found in the network. We determined this size by analyzing the observed signals obtained by applying the Fisher omnibus statistic to the list of candidate genes for different neighborhood sizes, and choosing the size for which we caught the best signal as the most reliable one. For three data sets (FXS, MFS, CF), we determined a small neighborhood because we observed best signal for small neighborhoods (data not shown). However, for the fourth data set (BMD), we determined a larger neighborhood because we could not observe a strong signal for smaller neighborhoods (data not shown). Therefore, we had to expand the number of neighboring genes that were taken into account for finding highly differentially expressed neighborhoods. This difference showed us that the size of the neighborhood can differ and is dependent on the number of affected genes in the disease and their biological pathways.

If we chose a neighborhood size for which we caught a weak signal, the ranking could produce an unreliable result. For example, $\beta = 0.5$ for FXS [11] that leads to a neighborhood size of 20 genes (see Table S9) would capture no meaningful signal from the neighbors and would lead to an unreliable ranking. Instead, a ranking for which a strong signal regarding to the significance is observed can be seen as a reliable result. Many factors influence the selection of an optimal neighborhood size: the shape of the weighting window, the density of the network (i.e., average node degree), the choice of the index of differential expression (here, the logarithm of the fold change), and so on. Further optimization of all the parameters will be necessary to devise an optimal procedure for neighborhood size selection.

Fourth, technical details of the mathematical model could have a substantial influence on the performance. Our purpose was to identify all direct and indirect connections in the functional protein association network from STRING that leads to a densely connected (global distance) network. We computed the Laplacian Exponential Diffusion Kernel [36] to obtain a global distance network. Following *Fouss et al. (2006)*, there exists several kernels on graphs, such as the Exponential Diffusion Kernel, the regularized Laplacian Kernel, the von Neumann Diffusion Kernel, and the Commute Time Kernel [37]. We chose the Laplacian Exponential Diffusion Kernel because *Köhler et al. (2008)* already applied this kernel as a good performing kernel to construct a distance network [3]. Furthermore *Fouss et al. (2006)* observed that this kernel performs systematically better than their corresponding adjacency matrix-based kernels (Exponential Diffusion Kernel, von Neumann Diffusion Kernel) [37]. However, a more systematic assessment on the effect of the kernel and its comparison to shortest-path, direct-interaction-only, or other methods, such as GTOM [38] is certainly in order. To efficiently compute the kernel matrix we applied the Cholesky Decomposition (see Materials and Methods) because the naïve computation was too time consuming. However, by calculating the Cholesky Decomposition we could compute the kernel method successfully with full accuracy (see Tables S7, S8). As a measure for determining the change in expression of a gene we chose the fold-change between control and experiment [39]. It is a commonly used measure (e.g., [11–14]), but numerous alternatives exist and will be compared in future work, such as differential expression indices expressed in terms of $z$-scores or log $p$-values, or combining the differential expression level and its statistical significance.

There are certainly diseases for which the candidate genes show no significances (i.e., there are no significant p-values that distinguish significant from not significant genes). In this case the disease does not lead to detectable disrupted expression modules or affected pathways in the network. Although it would be preferable to detect the disease gene, the ability of the method to return a negative result is an asset of the method. Indeed the performance is greatly influenced by the quality of the expression

data, the quality of the protein network around the actual disease gene, and the underlying biology (i.e., existence of actually disrupted expression modules) so that the method cannot be expected to work optimally on every disorder. Therefore, the ability to assess statistical significant and potentially return a negative result (i.e., "no strong candidate found") makes it possible to avoid pursuing "best" candidates that are actually not promising. There are also cases for which our method does return significant genes among which the known disease causing gene can however not be found. For those cases, one or several disrupted expression modules are found but none of them is centered on the disease-causing gene. This can be explained by the fact that even a single-gene disorder can result in the disruption of several pathways (downstream of the original pathway), making therefore the signal we are looking for more difficult to detect. This reflects that the current method is not suitable for all single gene disorders.

We applied our approach to Stein-Levental syndrome [22] and could identify two important candidate genes (FST and FBN3) from two different chromosomal regions [23–26]. We further detected the DEAD box 4 gene that is located on chr5q11.2 and that was ranked on the top position with a significant p-value. DEAD box polypeptide 4 (synonyms Mvh and Vasa) is an RNA helicase and is used commonly as a marker for germline cells [28,29]. Although little is know about the molecular function of DEAD box 4 in mammals, there are several reasons why Dead box 4 is a plausible candidate gene for Stein-Levental syndrome. First, in male mice loss of Dead box 4 results in infertility due to suspension of premeiotic differentiation of spermatogenic cells [30]. Female knockout mice do not show any obvious reproductive defects, but it is plausible that the effect may be less dramatic in females, leading only to a (partial) arrest in follicle development. One of the genes that fail to be expressed in a mouse Dead box 4 knockout is Aven, a caspase inhibitor [29]. Failure to express Aven thus may result in altered apoptosis control. Stein-Levental syndrome is characterized by follicular arrest, where several follicles develop to a size of 5–7 mm but not further. Since primary follicles secrete androgens, too many arrested follicles likely lead to elevated androgen levels, a hallmark of Stein-Levental syndrome. Second, DEAD box 4 interacts with Dicer1 and is colocalized with MILI [31]. Dicer1 processes miRNA precursors to mature miRNAs which are incorporated in the RISC complex. These miRNA-RISC complexes then exert a broad posttranscriptional control on many mRNAs. Mili belongs to the Piwi family, encodes a component of the RISC complex and is expressed at early stages of oocyte growth [32] and crucial for progression through spermatogenesis [33]. In Drosophila, the loss of Piwi function leads to the failure of germline cyst formation [34]. As a third potential mechanism, altered DEAD box 4 may influence stem cell recruitment to the ovaries. *Johnson et al.* have demonstrated that stem cells from bone marrow and peripheral blood which are also marked with the presence of DEAD box 4, can migrate to the ovaries of sterilized females and give rise to oocyte-containing follicles. Interestingly, the level of DEAD box 4 is influenced by the estrous cycle [35].

In summary, we have developed a novel gene prioritization approach that substitutes expression data to prior knowledge of the molecular basis of the disease, as required by existing methods. Our method ranks candidate genes by their differentially expressed neighborhoods. We have developed an efficient algorithm to compute a genomewide network by applying the Cholesky decomposition. To illustrate the power of our method, we have applied it to four constitutional genetic disorders and successfully prioritized the known disease-causing genes. In an application on a genetic disorder that is not yet well studied, we could retrieve two important known candidate genes and suggest a new candidate gene.

## Materials and Methods

### Overview

A network model based on the STRING database [8] was built for human genes. This sparse network model was used to compute distances between single genes via the Laplacian Exponential Diffusion Kernel [36]. The goal of such an approach is to build a network model based on global similarity measure (i.e., to include both direct and indirect connections between genes).

For each disease, a set of candidate genes (e.g., from a linkage study) has to be defined and a relevant microarray experiment is required (e.g., diseased sample vs. reference sample). A main issue that arises when applying kernel graph methods to genomewide networks is the computing time. To reduce it, the kernel matrices were approximated by the full or incomplete Cholesky decomposition and the reduced eigenvalue decomposition. Knowing the distances between single genes in the network, the differential expression level of adjacent genes can be considered. Each candidate gene receives a score based on the analysis of the differentially expressed neighborhoods, by which it is ranked. Candidate genes with genes in their neighborhood having highly differentially expressed indices are strong candidates.

### Functional protein association network

A protein association network is an undirected graph with proteins as nodes and weights as edges. If there exists an association between two proteins, an edge will be set between the corresponding nodes in the graph. The weights of the edges are taken represent the probability that such an association exists in reality. To model our functional protein association network, we used data from STRING [8] (version 7.1). STRING is a database of known and predicted protein-protein interactions. The interactions are derived from different information sources and different organisms, whereas the interactions include physical and functional associations. To build our networks, we have used the fused network provided by STRING that already integrates all available protein-protein interactions.

### Kernel matrix and distance network

We hypothesize that global network-similarity measures capture relationships between disease proteins better than algorithms based on direct interactions. To capture global relationships within a graph, a graph kernel was used. A graph kernel computes the global similarity of two nodes as the probability of reaching one node at some time point after a random walk starting from another node. This global similarity detects, besides direct, also indirect connections in the graph. The resulting graph leads to a global distance network where the edge between two nodes does not represent a direct interaction, but rather the global distance in this network.

The **Laplacian Exponential Diffusion Kernel** was introduced by *Kondor and Lafferty (2002)* [36] as

$$K = \lim_{n\to\infty}\left(I + \frac{\beta L}{n}\right)^n = e^{\beta L} \qquad (1)$$

whereby $L$ is the Laplacian matrix of a weighted and undirected graph $G$ with symmetric weights [36,37], and the parameter $\beta$ is the diffusion parameter that determines the degree of diffusion (for more details see Supplementary Materials). For a Laplacian matrix,

$e^{\beta L}$ is always positive definite and thus can be used as a kernel matrix. It can be seen as a random walk, starting from a node and transitioning to a neighboring node with the probability $\beta$.

The complexity of Equation (1 is $O(n^3)$. To improve the performance of the computation of such a diffusion, we can apply the Cholesky decomposition, that is decomposing the Laplacian matrix into the product of a lower triangular matrix by its transpose, before computing the kernel from this matrix. In this way, the computing time is reduced significantly as presented in Table S7.

The resulting kernel matrix is a densely connected network. This new network detects not only the direct interaction from the original protein association network, but also all indirect interactions via other genes. Thanks to this property the distances between all genes in the network can be determined. They are necessary for identifying highly expressed neighborhoods with a certain distance from a candidate gene, even if the genes are not directly interacting.

## Cholesky decomposition

There exists for every symmetric positive definite matrix $A$ of rank $n$ exactly one lower triangular matrix $S=(s_{ij})$ with $s_{ij}=0$ if $i<j$ and $s_{ij}>0$ if $i=1, 2,\ldots, n$.

The Cholesky decomposition (CD) solves the linear equation $A\cdot x=b$ by transforming $A$ into a product of a lower triangular matrix $S$ of rank $n$ and its transposed $S^T$:

$$A=S\cdot S^T \tag{2}$$

To reduce the dimensionalities of kernel matrices we applied the Incomplete Cholesky decomposition (ICD) with pivoting [40–42].

ICD is an iterative algorithm that approximates a lower rank matrix ($m\ll n$) in order to reduce the dimensionalities so that

$$A\approx\tilde{S}\cdot\tilde{S}^T,\in R^{n\times m},\ m\ll n \tag{3}$$

The resulting matrix $\tilde{S}$ is a lower triangular matrix of rank $m$. The overall complexity is $O(m^2n)$ and the storage requirement is $O(mn)$.

## Distance network

Having a symmetric and positive semidefinite Laplacian matrix L, we can compute the Laplacian Exponential Diffusion Kernel by applying CD or ICD instead of Equation (1), the computing time can be reduced significantly from $O(n^3)$ to $O(n^2m)$, in which $m$ is the rank of $S$ (Equations 2–3)).

In Table S7, we show that the computing time in calculating the Laplacian Exponential Diffusion Kernel can be significantly reduced by applying the CD. If we can accept an approximation to the exact kernel, we can apply the ICD by defining a threshold to reduce the rank of the matrices. In our example, we chose a threshold that led to an error of 7%–10%, depending on the size of the matrix (Table S8). The error is the difference of the norm of the reduced rank matrix and of the full rank matrix. For a network of the whole human genome containing 16,566 genes and an error of 7%, the resulting matrix got a low rank matrix with a rank of 1,829 that could be computed in 35 minutes, whereas by running the computation by applying the initial equation introduced by *Kondor and Lafferty (2002)* [36] (Equation (1)), the computation on the same machine (dual Opteron 250 with 16 GB RAM) ran out of memory and could not finish the computation. Even if no error can be handled and the CD is applied, and thus the matrices remain full rank, the computation could be successfully finished after 32 hours.

## Gene expression analysis

After obtaining a global distance network by computing the Laplacian Exponential Diffusion Kernel from the protein association network, the gene expression profiles for a specific disease can be mapped onto this network. For this purpose, relevant genome-wide microarray experiment data (e.g., disease sample vs. reference sample) is required from which the differential expression level for each gene in the genome is computed. Basing on the microarray experiment (the datasets that we chose from GEO were from good quality and already normalized) the fold change between the two conditions can be computed for each gene. We applied the fold-change between control and experiment [39]. For our method only the absolute value of the fold-change is relevant (i.e., if a gene is highly differentially expressed or not). We do not use a threshold to distinguish between highly and lowly differentially expressed genes, because our method considers all differential expression levels for computing the scores.

## Scoring and ranking

For each candidate gene a score is computed based on the differential expression levels of its neighborhood. For this purpose the differential expression levels of all neighbors in the distance network are ordered by their distance to the candidate gene. The rank of the diffusion distance is then taken as the new distance measure. The differential expression levels are then multiplied by a weighting function ($w=e^{-\beta\cdot r}$, controlled by parameter $\beta$ and rank $r$) to stress expression of close neighbors and to suppress expression of neighbors being far. Then the expression data is distributed randomly over the network. The score for a candidate gene is defined by the maximum deviation between its weighted neighboring expression and the randomized expression. Thus, the higher the level of differential expression level of close neighboring genes, the higher the score for the candidate gene (see Figure 3).

To obtain an empirical p-value for each candidate gene, the distribution of the scores was determined by repeatedly (3,000 times) randomly distributing the expression data over the network to estimate the significance of the signals of the actual candidates. By comparing the score of each candidate gene with the random distribution of the scores, a $p$-value was assigned. If a candidate's score was larger than 95% ($\alpha=0.05$) of all randomized scores, this candidate gene's score was considered significant.

## Neighborhood size

For our soft neighborhoods, we also define a notion of *neighborhood size*. Instead of having a hard threshold on the distance that determines whether a gene belongs to the neighborhood or not, our weighting function has a scale parameter that determines how quickly the weight decreases as a function of the distance. This parameter is what defines our neighborhood "size". To communicate about the neighborhood size easily, we use the following procedure. First, a threshold on the weight is selected, meaning that we neglect the contribution of all genes that have a weight lower than the threshold and are thus sufficiently "far away" from the candidate (the number of considered genes varies between 20 and 4000, depending on $\beta$ that varies from 0.001 and 0.5 (see Table S9)). As a result, we view the genes for which the weight is considered nonnegligible as "inside" the neighborhood. In our case, the weighting function is a negative exponential function ($w=e^{-\beta r}$) regulated by a parameter $\beta$. As $\beta$ decreases, the neighborhood size increases, and more neighboring genes are considered for computing the scores of the candidates (see Table S9).

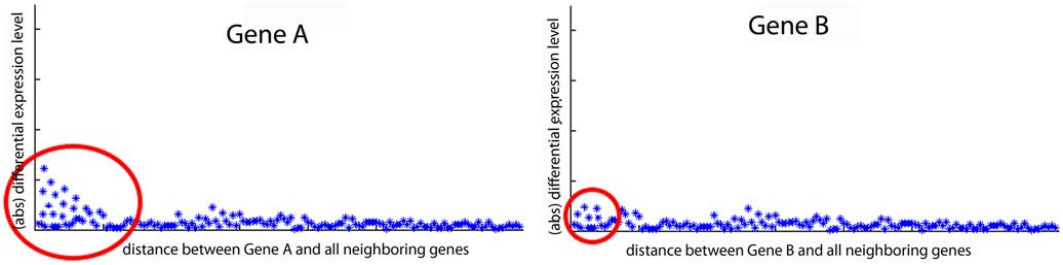



**Figure 3. Level of differential expression as function of distance from the candidate for two different candidate genes.** *Gene A* has a more differentially expressed genes among its close neighboring genes in the distance network than *gene B*, and is therefore a better candidate gene than *gene B*.
doi:10.1371/journal.pone.0005526.g003

For each data set, an appropriate neighborhood size must be computed independently. Therefore, we first run the analysis for various value of $\beta$ (from 0.001 to 0.5) and, for each, measure the signal captured by using the Fisher omnibus statistic ($S = \sum -2\ln(p-value)$) on the rankings produced. We then generate a new $p$-value from the statistic $S$ for each $\beta$ using the $\chi^2$ distribution. The value of the parameter $\beta$ with the smallest corresponding $p$-value is considered the appropriate neighborhood size for this data set. The idea is that, for an appropriate neighborhood size, some of the candidates will capture meaningful signal from their neighbors. By contrast, for inappropriate neighborhood sizes, all candidates will have uniformly distributed $p$-values, leading to a low statistic $S$.

Table S9 illustrates the signals derived by the Fisher omnibus meta-analysis using the example of FXS [11], leading to an appropriate neighborhood size of approximately 150 genes for this data set.

## Data sets

- FXS
  - Mendelian disorder: fragile X syndrome (FXS, OMIM #300624)
  - Disease gene: fragile X mental retardation 1 (FMR1, OMIM *309550)
  - Phenotype: mental retardation, macroorchidism, and distinct facial features
  - Expression data: *Nishimura et al. (2007)* [11] (GEO accession number: GSE7316). Lymphoblastoid cell cultures from patients with confirmed FMR1 full mutation (CGG repeat expansion). Platform: Agilent-012391 Whole Human Genome Oligo Microarray G4112A
- MFS
  - Mendelian disorder: Marfan syndrome (MFS, OMIM #154700)
  - Disease gene: fibrillin-1 precursor (FBN1, OMIM *134797)
  - Phenotype: variable skeletal abnormalities, tall stature, disproportionately long limbs and digits, joint laxity, eye anomalies and progressive cardiovascular problems.
  - Expression data: *Yao et al. (2007)* [12] (GEO accession number GDS2960). Fibroblast cultures from patients with confirmed FBN1 missense (9) and nonsense (7) mutations as well as one multi-exon deletion. Platform: Research Genetics (Invitrogen) - GF211 Microarray Filter
- CF
  - Mendelian disorder: Cystic fibrosis (CF, OMIM #219700)
  - Disease gene: Cystic fibrosis transmembrane conductance regulator (CFTR, OMIM *602421, [43])
  - Phenotype: chronic obstructive lung disease, bronchiectasia, and exocrine pancreatic insufficiency
  - Expression data: *Wright et al. (2006)* [13] (GEO accession number GDS2143). Analysis of the nasal respiratory epithelium of cystic fibrosis (CF) patients with mild (4) or severe (5) lung disease. Platform: Affymetrix GeneChip Human Genome U133 Array Set HG-U133B
- BMD
  - Mendelian disorder: Becker muscular dystrophy (BMD, OMIM #300376, [44])
  - Disease gene: dystrophin (DMD, OMIM *300377)
  - Phenotype: muscle wasting and weakness, and in some cases with mental impairment.
  - Expression data: *Bakay et al. (2006)* [14] (GEO accession number GDS2855). Analysis of muscle biopsy specimens from patients with various muscle diseases. Platform: Affymetrix GeneChip Human Genome U133 Array Set HG-U133B
- Stein-Levental syndrome
  - Mendelian disorder: Stein-Levental syndrome ( OMIM %184700)
  - Putative disease genes: follistatin (FST (HGNC symbol), OMIM *136470) putatively related with Stein-Levental syndrome [24], fibrillin 3 (FBN3), OMIM *608529) putatively associated with Stein-Levental syndrome [26].
  - Phenotype: obesity, hyperandrogenism and chronic anovulation
  - Expression data: *Cortón et al. (2007)* [22] (GEO accession number: GDS2084). Omental fat biopsy from patients. Unconfirmed disorder etiology. Platform: Affymetrix GeneChip Human Genome U133 Array Set HG-U133A

## Supporting Information

**Supplementary Materials S1**
Found at: doi:10.1371/journal.pone.0005526.s001 (0.07 MB DOC)

**Figure S1** Neighborhood of FMR1 and DUSP9. These graphs show the closest neighbors of FMR1 (A) and DUSP9 (B) including their differential expression levels and distances to FMR1 and DUSP9. For genes with absolute differential expression levels (2-fold-changes) larger than 1.5, the nodes are labeled with the gene names, the node size increases with the value and the color gets darker. With decreasing distances (i.e., increasing similarities) the edges between FMR1 or DUSP9 and their neighbors become thicker. This picture shows that the neighborhood of FMR1 belongs to a more disrupted expression module than the neighborhood of DUSP9.
Found at: doi:10.1371/journal.pone.0005526.s002 (2.11 MB TIF)

**Figure S2** Neighborhood of FBN1 and LEO1. These graphs show the closest neighbors of FBN1 (A) and LEO1 (B) including their differential expression levels and distances to FBN1 and LEO1. For genes with large absolute differential expression levels (2-fold-changes), the nodes are labeled with the gene names, the node size increases with the value and the color gets darker.With decreasing distances (i.e., increasing similarities) the edges between FBN1 or LEO1 and their neighbors become thicker. This picture shows that the neighborhood of FBN1 belongs to a more disrupted expression module than the neighborhood of LEO1.
Found at: doi:10.1371/journal.pone.0005526.s003 (2.43 MB TIF)

**Figure S3** Neighborhood of CFTR and PIK3CG. These graphs show the closest neighbors of CFTR (A) and PIK3CG (B) including their differential expression levels and distances to CFTR and PIK3CG. For genes with large absolute differential expression levels (2-fold-changes), the nodes are labeled with the gene names, the node size increases with the value and the color gets darker. With decreasing distances (i.e., increasing similarities) the edges between CFTR or PIK3CG and their neighbors become thicker. This picture shows that the neighborhood of CFTR belongs to a more disrupted expression module than the neighborhood of PIK3CG.
Found at: doi:10.1371/journal.pone.0005526.s004 (2.22 MB TIF)

**Table S1** Top 25 ranked candidate genes in Fragile X syndrome (FXS). Fragile X syndrome [11] is a disorder caused by mutation in the FMR1 gene, and is characterized by mental retardation, macroorchidism, and distinct facial features. Candidate genes were chosen from chrXq26-q28 that contains 119 genes including FMR1. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas the top eleven genes have significant p-values ($\alpha = 0.05$). FMR1 ranked first with a significant p-value (0.00131), and FMR2, also involved in FXS, got a significant p-value of 0.01991 on position 6. Out of the eleven significant candidate genes in the ranking we identified four genes, including FMR1 and FMR2, that are known to be linked to mental retardation [11,15–17].
Found at: doi:10.1371/journal.pone.0005526.s005 (0.06 MB DOC)

**Table S2** Top 25 ranked candidate genes in Marfan syndrome (MFS). Marfan syndrome [12] is a heritable connective tissue disorder caused by mutations in the FBN1 gene, and is characterized by increased height, disproportionately long limbs and digits, anterior chest deformity, joint laxity, vertebral column deformity, and other variable skeletal abnormalities, as well as several ocular and cardiovascular features. Candidate genes were chosen from 15q15.3-q22.33 that contains 129 genes including FBN1. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas the top six genes have significant p-values ($\alpha = 0.05$). FBN1 was ranked on the fifth position with a significant p-value (0.0226). In the ranking we obtained six genes that were significant but not involved in MFS or phenotype related diseases.
Found at: doi:10.1371/journal.pone.0005526.s006 (0.06 MB DOC)

**Table S3** Top 25 ranked candidate genes in Cystic fibrosis (CF). Cystic fibrosis [13] is an autosomal recessive disorder of epithelial ion transport caused by mutations in the CF transmembrane conductance regulator gene (CFTR), and is characterized by chronic obstructive lung disease, bronchiectasia, and exocrine pancreatic insufficiency. Candidate genes were chosen from chr7q22.1-31.33 that contains 110 genes including CFTR. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas the top nine genes have a significant p-value ($\alpha = 0.05$). CFTR was ranked in the seventh position with a significant p-value (0.046). In the ranking we obtained seven genes that were significant but not involved in CF or phenotype related diseases. However, out of the top 25 ranked genes we detected four genes that are known to be linked to CF [18–21].
Found at: doi:10.1371/journal.pone.0005526.s007 (0.06 MB DOC)

**Table S4** Top 25 ranked candidate genes in Becker muscular dystrophy (BDM). Becker muscular dystrophy [14] is a X-linked progressive myopathy caused by mutations within the DMD gene, and is characterized by muscle wasting and weakness, and in some cases with mental impairment. Candidate genes were chosen from chrXp22.33-21.1 that contains 116 genes including DMD. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas the top two genes have significant p-values ($\alpha = 0.05$). DMD ranked on the second position with a significant p-value (0.0272). The other significant gene is not involved in BMD or in a phenotype related disease.
Found at: doi:10.1371/journal.pone.0005526.s008 (0.06 MB DOC)

**Table S5** Top 25 ranked candidate genes from chr5q11.2 in Stein-Levental syndrome. Stein-Levental syndrome [22] is a oligogenic hormonal disorder among women putatively related with the FST gene, and is characterized by hyperandrogenism, chronic anovulation and associated with obesity. Candidate genes were chosen from chr5q11.2 that contains 25 genes including the candidate gene FST [24]. These candidate genes were ranked by our new approach, whereas only the top gene has a significant p-value ($\alpha = 0.05$). FST was ranked in the second position with a p-value of 0.056, and we received only one significant gene (DEAD box 4) that is a plausible candidate gene for Stein-Levental syndrome.
Found at: doi:10.1371/journal.pone.0005526.s009 (0.06 MB DOC)

**Table S6** Top 25 ranked candidate genes from chr19p13.2 in Stein-Levental syndrome. Stein-Levental syndrome [22] is a oligogenic hormonal disorder among women putatively associated with the FBN3 gene, and is characterized by hyperandrogenism, chronic anovulation and associated with obesity. Candidate genes were chosen from chr19p13.2 that contains 100 genes including the candidate gene FBN3 [26]. These candidate genes were ranked by our new approach, and the top 25 ranked candidate

genes are presented here, whereas only the top gene has a significant p-value ($\alpha = 0.05$). FBN3 was ranked in the fifth position with a p-value of 0.0595.
Found at: doi:10.1371/journal.pone.0005526.s010 (0.06 MB DOC)

**Table S7** CPU times for computing the Laplacian Exponential Diffusion Kernel by its definition, the CD and the ICD. CPU times (sec) for computing the Laplacian Exponential Diffusion Kernel by its definition (Equation 1), by CD (Equation 2), and by ICD (Equation 3) with a threshold leading to an error of 7%–10%. The computation were run on a dual Opteron 250 with 16 GB RAM.
Found at: doi:10.1371/journal.pone.0005526.s011 (0.03 MB DOC)

**Table S8** Rank and Error of resulting Laplacian Exponential Diffusion Kernel computed by its definition, the CD and the ICD. Rank and Error of resulting Laplacian Exponential Diffusion Kernel computed by its definition, (Equation 1), by CD (Equation 2), and by ICD (Equation 3) with a threshold leading to an error of 7%–10%. The computation were run on a dual Opteron 250 with 16 GB RAM.
Found at: doi:10.1371/journal.pone.0005526.s012 (0.03 MB DOC)

**Table S9** Example for determining an appropriate neighborhood size using the example of data set 1 (FXS [11]). The neighborhood size is controlled by a weighting function ($w = \exp(-\beta \cdot r)$). Applying the Fisher omnibus meta-analysis ($S = \sum -2 \ln$ (p-value)) for each parameter $\beta$, new p-values are generated from a $X/\backslash 2$ distribution. The parameter $\beta$ , for which the smallest p-value is observed (here: $\beta = 0.05$), leads to the appropriate neighborhood size for FXS (approx. 150 genes).
Found at: doi:10.1371/journal.pone.0005526.s013 (0.02 MB DOC)

## Acknowledgments

The authors would like to thank Marco Saerens from Université Catholique de Louvain (UCL) for helpful comments on the Cholesky Decomposition.

## Author Contributions

Analyzed the data: DN BT LT HVE KD. Contributed reagents/materials/analysis tools: DN LCT. Wrote the paper: DN. Supervision: YM.

## References


1. Bubendorf L (2001) High-throughput microarray technologies: from genomics to clinics. Eur Urol 40(2): 231–8.
2. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. Nat Biotech 24(5): 537–44.
3. Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82(4): 949–58.
4. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78(6): 1011–25.
5. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature Biotechnol 25(3): 309–16.
6. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140.
7. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38(3): 285–93.
8. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7 - recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358–62.
9. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. (2007) ArrayExpress-a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35(Database issue): D747–50.
10. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2007) NCBI GEO: mining tens of millions of expression profiles - database and tools update. Nucleic Acids Res 35(Database issue): D760–5.
11. Nishimura Y, Martin CL, Vazquez-Lopez A, Spence SJ, Alvarez-Retuerto AI, et al. (2007) Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. Hum Mol Genet 16(14): 1682–98.
12. Yao Z, Jaeger JC, Ruzzo WL, Morale CZ, Emond M, et al. (2007) A Marfan syndrome gene expression phenotype in cultured skin fibroblasts. BMC Genomics 8: 319.
13. Wright JM, Merlo CA, Reynolds JB, Zeitlin PL, Garcia JGN, et al. (2006) Respiratory epithelial gene expression in patients with mild and severe cystic fibrosis lung disease. Am J Respir Cell Mol Biol 35(3): 327–36.
14. Bakay M, Wang Z, Melcon G, Schiltz L, Xuan J, et al. (2006) Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration. Brain 129(Pt 4): 996–1013.
15. Probst FJ, Roeder ER, Enciso VB, Ou Z, Cooper ML, et al. (2007) Chromosomal microarray analysis (CMA) detects a large X chromosome deletion including FMR1, FMR2, and IDS in a female patient with mental retardation. Am J Med Genet 143A(12): 1358–65.
16. Rosenberg EH, Almeida LS, Kleefstra T, deGrauw RS, Yntema HG, et al. (2004) High prevalence of SLC6A8 deficiency in X-linked mental retardation. Am J Hum Genet 75(1): 97–105.
17. Moretti P, Zoghbi HY (2006) MeCP2 dysfunction in Rett syndrome and related disorders. Curr Opin Genet Dev 16(3): 276–81.
18. Mount DB, Romero MF (2004) The SLC26 gene family of multifunctional anion exchangers. Pflugers Arch 447(5): 710–21.
19. Moehle C, Ackermann N, Langmann T, Aslanidis C, Kel A, et al. (2006) Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. J Mol Med 84(12): 1055–66.
20. Malmberg EK, Noaksson KA, Phillipson M, Johansson ME, Hinojosa-Kurtzberg M, et al. (2006) Increased levels of mucins in the cystic fibrosis mouse small intestine, and modulator effects of the Muc1 mucin expression. Am J Physiol Gastrointest Liver Physiol 291(2): G203–10.
21. Kyo K, Muto T, Nagawa H, Lathrop GM, Nakamura Y (2006) Associations of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease. J Hum Genet 46(1): 5–20.
22. Cortón M, Botella-Carretero JI, Benguría A, Villuendas G, Zaballos A, et al. (2007) Differential gene expression profile in omental adipose tissue in women with polycystic ovary syndrome. J Clin Endocrinol Metab 92(1): 328–37.
23. Franks S (2006) Candidate genes in women with polycystic ovary syndrome. Fertil Steril 86(Suppl 1): S15.
24. Jones MR, Wilson SG, Mullin BH, Mead R, Watts GF, et al. (2007) Polymorphism of the follistatin gene in polycystic ovary syndrome. Mol Hum Reprod 13(4): 237–41.
25. Urbanek M, Woodroffe A, Ewens KG, Diamanti-Kandarakis E, Legro RS, et al. (2005) Candidate gene region for polycystic ovary syndrome on chromosome 19p13.2. J Clin Endocrinol Metab 90(12): 6623–9.
26. Urbanek M, Sam S, Legro RS, Dunaif A (2007) Identification of a polycystic ovary syndrome susceptibility variant in fibrillin-3 and association with a metabolic phenotype. J Clin Endocrinol Metab 92(11): 4191–8.
27. Simoni M, Tempfer CB, Destenaves B, Fauser BCJM (2008) Functional genetic polymorphisms and female reproductive disorders: Part I: polycystic ovary syndrome and ovarian response. Hum Reprod Update 14(5): 459–484.
28. Noce T, Okamoto-Ito S, Tsunekawa N (2001) Vasa homolog genes in mammalian germ cell development. Cell Struct. Funct 26: 131–136.
29. Ina S, Tsunekawa N, Nakamura A, Noce T (2003) Expression of the mouse Aven gene during spermatogenesis, analyzed by subtraction screening using Mvh-knockout mice. Gene Expr Patterns 3: 635–638.
30. Tanaka SS, Toyooka Y, Akasu R, Katoh-Fukui Y, Nakahara Y, et al. (2000) The mouse homolog of Drosophila Vasa is required for the development of male germ cells. Genes Dev 14: 841–853.
31. Kotaja N, Bhattacharyya SN, Jaskiewicz L, Kimmins S, Parvinen M, et al. (2006) The chromatoid body of male germ cells: similarity with processing bodies and presence of Dicer and microRNA pathway components. Proc Natl Acad Sci U S A 103: 2647–2652.
32. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, et al. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature 453: 539–543.
33. Kuramochi-Miyagawa S, Kimura T, Ijiri TW, Isobe T, Asada N, et al. (2004) Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. Development 131: 839–849.

# Supplementary Material

## Case Studies

### Case Study 1: Fragile X syndrome (FXS)

The first data set we ran with our method was a microarray analysis to compare the mRNA expression profile in lymphoblastoid cells from patients with autism due to a fragile X mutation (FMR1-FM) and non-autistic controls [11]. Fragile X syndrome (FXS) is a disorder caused by expansion of the CGG trinucleotide repeat in the 5' untranslated region of the fragile X mental retardation 1 (FMR1) gene gene located on chrXq27.3, and is characterized by mental retardation, macroorchidism, and distinct facial features.

Candidate genes were chosen from chrXq26-q28 that contains 119 genes including FMR1. The top 25 ranked genes are presented in Supplementary Table 1.

**Neighborhood**

For the ranking of FXS a neighborhood of 150 genes was determined, i.e. we have considered the 150 genes with the smallest distances to the candidate genes, because larger neighborhoods were found to be meaningless for this ranking due to large changes in the expression for very close neighboring genes (data not shown).

Comparing the disrupted expression neighborhoods of FMR1 and the low ranked gene DUSP9 (see Supplementary Table 1), a graph containing up to the nearest 100 genes could be drawn (see Supplementary Figure 1). This visualization shows that the closest neighborhood of FMR1 is higher differentially expressed than of DUSP9, and that the neighborhood of FMR1 belongs to a more disrupted expression module than the neighborhood of DUSP9.

**Significance**

Figure 2-A shows the distribution of p-values after 10,000 randomizations, and the p-value of FMR1 which is significant. This figure illustrates that there are only few genes having significant p-values after a randomization procedure, and that the high rank of FMR1 is not random.

**Ranking**

FMR1 ranked first with a significant p-value (0.001), and FMR2, also involved in mental retardation, got a significant p-value of 0.02 on rank 6 out of 119 ranked genes.

Out of the eleven significant candidate genes in the ranking we identified four genes, including FMR1 and FMR2 [16] that are known to be linked to mental retardation.

Mutation in the creatine transporter gene SLC6A8 causes the X-linked Creatine Deficiency syndrome, a mental retardation disease [17].

Mutations in the MECP2 gene cause Rett syndrome, a severe neurodevelopmental disorder including mental retardation (e.g. [18]).

Summarizing, four of the eleven significant genes in our ranking are directly associated with mental retardation.

## Case Study 2: Marfan syndrome (MFS)

The second data set we ran with our method was a microarray analysis to compare the mRNA expression profile in cultured skin fibroblasts from patients with Marfan syndrome (MFS) and controls [12]. MFS is a heritable autosomal dominant connective tissue disorder caused by mutations in the FBN1 gene, and is characterized by variable skeletal abnormalities, tall stature, disproportionately long limbs and digits, joint laxity, eye anomalies and progressive cardiovascular problems.

Candidate genes were chosen from 15q15.3-q22.33 that contains 129 genes including FBN1. Their top 25 ranked genes are presented in Supplementary Table 2.

**Neighborhood**

For the ranking of MFS a neighborhood of again 150 genes was determined, because the largest disruptions in expressed neighborhoods could be again observed in small neighborhoods (data not shown).

Comparing the disrupted expression neighborhoods of FBN1 and the low ranked gene LEO1 (see Supplementary Table 2), a graph containing up to the nearest 100 genes could be drawn (see Supplementary Figure 2). This visualization shows that the neighborhood of FBN1 belongs to a more disrupted expression module than the neighborhood of LEO1 because of further highly differentially expressed genes in its neighborhood.

**Significance**

Figure 2-B shows the distribution of p-values after 10,000 Randomizations, and the p-value of FBN1 which is significant. This figure illustrates that there are only few genes having significant p-values after a randomization procedure, and that the high rank of FBN1 is not random.

**Ranking**

FBN1 was ranked in the fifth position out of 129 ranked candidate genes with a significant p-value (0.02). In the ranking we obtained six genes that were significant but not involved in MFS or phenotype related diseases.

## Case Study 3: Cystic fibrosis (CF)

The third data set we ran with our new method was a high-density oligonucleotide microarray analysis of nasal respiratory epithelium tissues of patients with cystic fibrosis (CF) [13**]**. Cystic fibrosis is an autosomal recessive disorder of epithelial ion transport caused by mutations in the CF transmembrane conductance regulator gene (CFTR), and is characterized by chronic obstructive lung disease, bronchiectasia, and exocrine pancreatic insufficiency.

Candidate genes were chosen from chr7q22.1-31.33 that contains 110 genes including CFTR. Their ranking is presented in Supplementary Table 3.

**Neighborhood**

For the ranking of CF a neighborhood of 20 genes was determined, because differences in the expression could be already detected for small neighborhoods for the top ranked genes (data not shown), and an expansion of the neighborhood was not necessary.

Comparing the disrupted expression neighborhoods of CFTR and the low ranked gene PIK3CG (see Supplementary Table 3), a graph containing 100 neighboring genes could be drawn (see Supplementary Figure 3). This visualization shows that the higher ranked gene CFTR belongs to a more disrupted expression module than PIK3CG.

**Significance**

Figure 2-C shows the distribution of p-values after 10,000 Randomizations, and the p-value of CFTR which is significant and ranked on the seventh position. This figure illustrates that there are only few genes having significant p-values after a randomization procedure, and that the high rank of CFTR is not random.

**Ranking**

CFTR was ranked in the seventh position out of 110 ranked candidate genes with a significant p-value of 0.04. In the ranking we obtained seven genes that were significant but not involved in CF or phenotype related diseases. However, out of the top 25 ranked genes we detected four genes that are known to be linked to CF.

The SLC26 gene family, including the top ranked genes SLC26A4 and SLC26A5, encodes anion exchangers. SLC26 Cl(-)-HCO(3)(-) and Cl(-)-OH(-) exchange is activated by the CFTR gene implicating defective regulation of these exchangers in the reduced HCO(3)(-) transport seen in CF [19].

By microarray analysis, [20] found coordinated downregulation of mucins, including MUC17, in colon of patients with inflammatory bowel disease, including Crohn disease and ulcerative colitis, compared with controls. [21] found an increased expression of Muc3 (orthologue of human MUC17) in a in the small intestine of CF mice.

[22] reported an association of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease.

Summarizing, four of the top 25 ranked genes are directly associated with CF or a disease related phenotype.

## Case Study 4: Becker muscular dystrophy (BMD)

The fourth data set we ran with our new method was a microarray analysis of muscle biopsy specimens from patients with various muscle diseases (including BMD) [14]. Becker muscular dystrophy is a X-linked progressive myopathy caused by mutations within the DMD gene, and is characterized by muscle wasting and weakness, and in some cases with mental impairment.

Candidate genes were chosen from chrXp22.33-21.1 that contains 116 genes including DMD. The top 25 ranked genes are presented in Supplementary Table 4.

**Neighborhood**

For the ranking of BMD a neighborhood of 2000 genes was determined, because large differences in the expression could be detected only for large neighborhoods for the top ranked genes, whereas for smaller neighborhoods the top ranked genes had no significant difference, thus, the neighborhood had to be expanded (data not shown).

**Significance**

Figure 2-D shows the distribution of p-values after 10,000 Randomizations, and the p-value of DMD which is significant and ranked on the second position. This figure illustrates that there are only few genes having significant p-values after a randomization procedure, and that the high rank of CFTR is not random.

**Ranking**

DMD ranked on the second position out of 116 ranked candidate genes with a significant p-value (0.02). In the ranking we obtained two genes that were significant, including DMD. The other significant gene is not involved in BMD or in a phenotype related disease.

## Case Study 5: Stein-Leventhal syndrome

The fifth data set we ran with our new method was a microarray analysis of omental adipose tissues of morbidly obese patients with Stein-Leventhal syndrome [23]. Stein-Leventhal syndrome is a hormonal disorder among women characterized by obesity, hyperandrogenism and chronic anovulation, and is putatively associated with mutation of FST (follistatin)) [25] or FBN3 (fibrillin 3) [27].

Candidate genes were chosen from chr5q11.2 that contains 25 genes including FST, and from chr19p13.2 that contains 100 genes including FBN3. Their rankings are presented in Supplementary Tables 5-6.

**Neighborhood**

For the ranking of Stein-Leventhal syndrome a neighborhood of 2000 genes was determined, because large differences in the expression could be detected only for large neighborhoods for the top ranked genes, whereas for smaller neighborhoods the top ranked genes had no significant difference, thus, the neighborhood had to be expanded (data not shown).

**Ranking**

FST was ranked in the second position out of 25 candidate genes with a p-value of 0.056 (Supplementary Table 5), and we received only one significant gene (DEAD box 4) that is a plausible candidate gene for Stein-Leventhal syndrome.

FBN3 was ranked on the fifth position out of 100 candidate genes with a p-value of 0.0595 (Supplementary Table 6).

## Kernel matrix – Distance network

Consider a given weighted and undirected graph *G* with symmetric weights $w_{ij} > 0$ between all linked couple of nodes *i* and *j*. The weight $w_{ij} > 0$ increases with the importance of the relation between nodes *i* and *j*: the larger its value, the easier the communication through the edge. The Adjacency matrix *A* is defined as $a_{ij} = w_{ij}$ if the nodes *i* and *j* are directly connected and $a_{ij} = 0$ otherwise. The Laplacian matrix *L* of *G* is defined as $L = Diag(a_{i.}) - A$, with the diagonal degree matrix $Diag(a_{i.})$ whereby its entries are defined as $d_{ii} = a_{i.} = \sum_{j=1}^{n} a_{ij}$ [37,38].

The **Laplacian Exponential Diffusion Kernel** was introduced by *Kondor and Lafferty (2002)* [37] as

$$K = \lim_{n\to\infty}\left( I + \frac{\beta L}{n} \right)^n = e^{\beta L} \tag{1}$$

whereby the parameter $\beta$ is the diffusion parameter that determines the degree of diffusion. For a Laplacian matrix, $e^{\beta L}$ is always positive definite and thus can be used as a kernel matrix. The diffusion effect of the kernel can be illustrated by

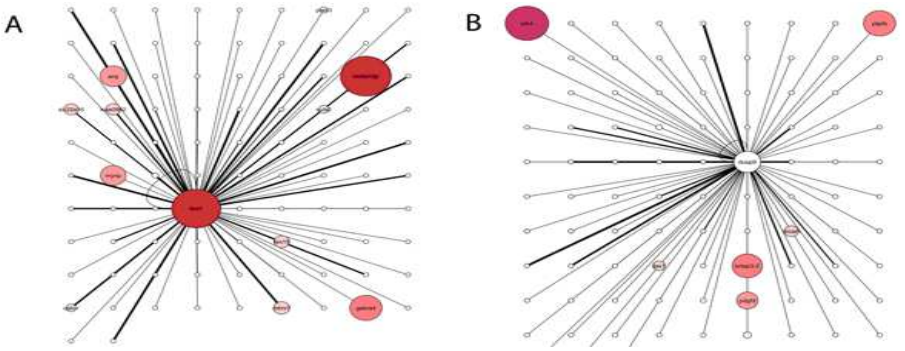
$$e^{\beta L} = I + \sum_{n=1}^{\infty}\left( \frac{\beta^n}{n!} L^n \right) \tag{2}$$

The kernel matrix *K* can be seen as a random walk, starting from a node and transitioning to a neighboring node with the probability $\beta$. Fixing the transition probability to every neighboring node as $\beta$, the probability of staying at the current node *i* is $1 - d_i\beta$, where $d_i$ is the degree of node *i*. The column vector *j* of *K* then represents the steady-state probability vector of the random walk when starting at node *j*, and the value $K_{ij}$ represents the probability that a random walk starting from *i* will be at *j* after infinite time steps. If the diffusion parameter is small, *K* can be seen as a lazy random walk because the probability that the random walker stays at the current node is high. As $\beta$ increases, the kernel values diffuse more completely through the graph, and when $\beta$ is sufficiently large, the values among distant nodes capture the long-range relationships between nodes [37].

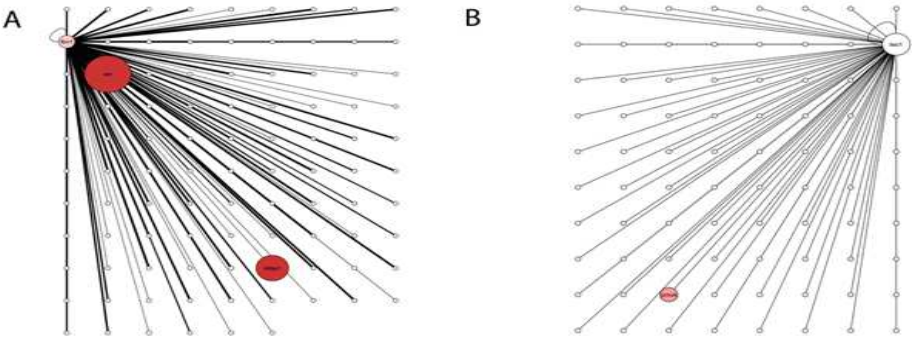
Beside the direct protein-protein interaction, the connectivity of the starting and ending node in a random walk play a central role in determining the global similarity in an interaction network.

The computation of a kernel matrix after Equation (1) is very time intensive for large networks, and its complexity is $O(n^3)$. If this kernel is computed for a whole genome, the computing time and the storage requirement must be improved. This can be reached by applying the Cholesky decomposition by decomposing the Laplacian matrix into a lower triangular matrix before computing the kernel from this matrix. By doing so, the computing time could be reduced significantly as presented in Supplementary Table 7.

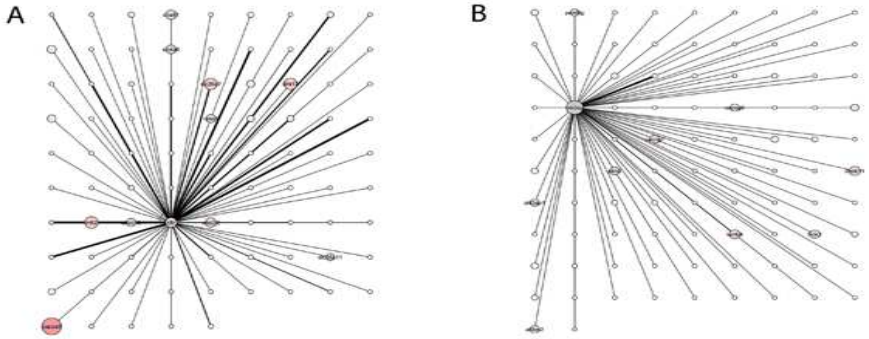
**Supplementary Figure S1: Neighborhood of FMR1 and DUSP9.**
These graphs show the closest neighbors of FMR1 (A) and DUSP9 (B) including their differential expression levels and distances to FMR1 and DUSP9. For genes with absolute differential expression levels (2-fold-changes) larger than 1.5, the nodes are labeled with the gene names, the node size increases with the value and the color gets darker. With decreasing distances (i.e., increasing similarities) the edges between FMR1 or DUSP9 and their neighbors become thicker. This picture shows that the neighborhood of FMR1 belongs to a more disrupted expression module than the neighborhood of DUSP9.




**Supplementary Figure S2: Neighborhood of FBN1 and LEO1.**
These graphs show the closest neighbors of FBN1 (A) and LEO1 (B) including their differential expression levels and distances to FBN1 and LEO1. For genes with large absolute differential expression levels (2-fold-changes), the nodes are labeled with the gene names, the node size increases with the value and the color gets darker.With decreasing distances (i.e., increasing similarities) the edges between FBN1 or LEO1 and their neighbors become thicker. This picture shows that the neighborhood of FBN1 belongs to a more disrupted expression module than the neighborhood of LEO1.



**Supplementary Figure S3: Neighborhood of CFTR and PIK3CG.**
These graphs show the closest neighbors of CFTR (A) and PIK3CG (B) including their differential expression levels and distances to CFTR and PIK3CG. For genes with large absolute differential expression levels (2-fold-changes), the nodes are labeled with the gene names, the node size increases with the value and the color gets darker. With decreasing distances (i.e., increasing similarities) the edges between CFTR or PIK3CG and their neighbors become thicker. This picture shows that the neighborhood of CFTR belongs to a more disrupted expression module than the neighborhood of PIK3CG.



**Supplementary Table S1: Top 25 ranked candidate genes in Fragile X syndrome (FXS).**

| Rank | Symbol | Score | 2fold-change | p-value | Linkage to phenotype |
|---|---|---|---|---|---|
| **1** | **FMR1** | **0.0028** | **4.81** | **0.0013** | **FXS [11,16]** |
| 2 | FAM11A | 0.0023 | 0 | 0.0036 | |
| 3 | SLC6A8 | 0.0016 | 0.13 | 0.0105 | X-linked mental retardation [17] |
| 4 | MAGEA9 | 0.0015 | 2.44 | 0.0164 | |
| 5 | MPP1 | 0.0015 | 2.75 | 0.0169 | |
| **6** | **FMR2** | **0.0014** | **1.08** | **0.0199** | **FXS** [16] |
| 7 | G6PD | 0.0013 | 0.60 | 0.0289 | |
| 8 | BRS3 | 0.0012 | 0.35 | 0.0365 | |
| 9 | SPANXA2 | 0.0012 | 0 | 0.0419 | |
| 10 | H2AFB2 | 0.0012 | 0.35 | 0.0421 | |
| 11 | MECP2 | 0.0011 | 1.10 | 0.0479 | Rett Syndrome [18] |
| 12 | ZNF185 | 0.0010 | 2.71 | 0.0699 | |
| 13 | MAGEC2 | 0.0010 | 1.09 | 0.0699 | |
| 14 | GAB3 | 0.0010 | 1.12 | 0.0721 | |
| 15 | TEX28 | 0.0010 | 0 | 0.0723 | |
| 16 | PLAC1 | 0.0010 | 0.40 | 0.0740 | |
| 17 | TFDP3 | 0.0010 | 1.39 | 0.0797 | |
| 18 | OPN1MW2 | 0.0010 | 0 | 0.0840 | |
| 19 | AVPR2 | 0.0009 | 0.13 | 0.1016 | |
| 20 | DUSP9 | 0.0009 | 0.60 | 0.1136 | |
| 21 | MAGEC1 | 0.0009 | 0.86 | 0.1158 | |
| 22 | IDS | 0.0008 | 0.63 | 0.1386 | |
| 23 | FHL1 | 0.0008 | 0.25 | 0.1399 | |
| 24 | MAGEA10 | 0.0008 | 1.58 | 0.1439 | |
| 25 | F8 | 0.0008 | 0.46 | 0.1440 | |

Fragile X syndrome [11] is a disorder caused by mutation in the FMR1 gene, and is characterized by mental retardation, macroorchidism, and distinct facial features.

Candidate genes were chosen from chrXq26-q28 that contains 119 genes including FMR1. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas the top eleven genes have significant p-values ($\alpha$=0.05). FMR1 ranked first with a significant p-value (0.00131), and FMR2, also involved in FXS, got a significant p-value of 0.01991 on position 6. Out of the eleven significant candidate genes in the ranking we identified four genes, including FMR1 and FMR2, that are known to be linked to mental retardation [11,16-18].

**Supplementary Table S2: Top 25 ranked candidate genes in Marfan syndrome (MFS).**

| Rank | Symbol | Score | 2fold-change | p-value | Linkage to phenotype |
|---|---|---|---|---|---|
| 1 | MFAP1 | 0.0152 | 0.15 | 0.0022 | |
| 2 | CYP19A1 | 0.0139 | 0 | 0.0042 | |
| 3 | FOXB1 | 0.0111 | 0 | 0.0151 | |
| 4 | PYGO1 | 0.0105 | 0 | 0.0197 | |
| **5** | **FBN1** | **0.0103** | 0.22 | **0.0226** | **MFS [12]** |
| 6 | DMXL2 | 0.0087 | 0 | 0.0417 | |
| 7 | ELL3 | 0.0076 | 0 | 0.0604 | |
| 8 | SMAD6 | 0.0075 | 0 | 0.0617 | |
| 9 | RAB8B | 0.0071 | 0 | 0.0718 | |
| 10 | CGNL1 | 0.0070 | 0 | 0.0725 | |
| 11 | TMOD3 | 0.0070 | 0 | 0.0731 | |
| 12 | FGF7 | 0.0067 | 0.14 | 0.0794 | |
| 13 | LCTL | 0.0067 | 0 | 0.0930 | |
| 14 | GRINL1A | 0.0065 | 0 | 0.1008 | |
| 15 | DAPK2 | 0.0065 | 0 | 0.1015 | |
| 16 | HERC1 | 0.0063 | 0 | 0.1091 | |
| 17 | PTPLAD1 | 0.0059 | 0.25 | 0.1171 | |
| 18 | NARG2 | 0.0057 | 0 | 0.1219 | |
| 19 | MYO5A | 0.0056 | 0.03 | 0.1268 | |
| 20 | RAB27A | 0.0053 | 0 | 0.1354 | |
| 21 | ALDH1A2 | 0.0049 | 0 | 0.1505 | |
| 22 | RAB11A | 0.0048 | 0.10 | 0.1539 | |
| 23 | MAP1A | 0.0045 | 0 | 0.1678 | |
| 24 | SPG21 | 0.0044 | 0 | 0.1785 | |
| 25 | TCF12 | 0.0041 | 0.03 | 0.1939 | |

Marfan syndrome [12] is a heritable connective tissue disorder caused by mutations in the FBN1 gene, and is characterized by increased height, disproportionately long limbs and digits, anterior chest deformity, joint laxity, vertebral column deformity, and other variable skeletal abnormalities, as well as several ocular and cardiovascular features.
Candidate genes were chosen from 15q15.3-q22.33 that contains 129 genes including FBN1. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas the top six genes have significant p-values ($\alpha$=0.05). FBN1 was ranked on the fifth position with a significant p-value (0.0226). In the ranking we obtained six genes that were significant but not involved in MFS or phenotype related diseases.

**Supplementary Table S3: Top 25 ranked candidate genes in Cystic fibrosis (CF).**

| Rank | Symbol | Score | 2fold-change | p-value | Linkage to phenotype |
|---|---|---|---|---|---|
| 1 | TFEC | 0.0010 | 1.32 | 0.0073 | |
| 2 | CAPZA2 | 0.0008 | 0 | 0.0117 | |
| 3 | WNT2 | 0.0008 | 0 | 0.0118 | |
| 4 | TAS2R16 | 0.0005 | 0 | 0.0270 | |
| 5 | MDFIC | 0.0005 | 0.98 | 0.0277 | |
| 6 | CAV2 | 0.0004 | 0 | 0.0380 | |
| **7** | **CFTR** | **0.0004** | **0.85** | **0.0460** | **CF[13]** |
| 8 | CTTNBP2 | 0.0003 | 0.82 | 0.0528 | |
| 9 | LAMB4 | 0.0003 | 0.82 | 0.0536 | |
| 10 | GNB2 | 0.0003 | 0 | 0.0574 | |
| 11 | ZNF3 | 0.0003 | 0.81 | 0.0582 | |
| 12 | SLC13A1 | 0.0002 | 0 | 0.0801 | |
| 13 | SLC26A5 | 0.0002 | 0 | 0.0802 | [19] |
| 14 | SLC26A4 | 0.0002 | 0 | 0.0803 | [19] |
| 15 | SMURF1 | 0.0002 | 0.73 | 0.0854 | |
| 16 | ACTL6B | 0.0002 | 0 | 0.1083 | |
| 17 | MUC17 | 0.0002 | 0.39 | 0.1280 | [20,21] |
| 18 | ZNF277P | 0.0002 | 0 | 0.1330 | |
| 19 | PUS7 | 0.0002 | 0 | 0.1344 | |
| 20 | THAP5 | 0.0002 | 0.64 | 0.1349 | |
| 21 | MUC3B\|MUC3A | 0.0001 | 0 | 0.1477 | [22] |
| 22 | DOCK4 | 0.0001 | 0.62 | 0.1541 | |
| 23 | CADPS2 | 0.0001 | 0 | 0.1546 | |
| 24 | IMMP2L | 0.0001 | 0 | 0.1561 | |
| 25 | PIK3CG | 0.0001 | 0 | 0.1652 | |

Cystic fibrosis [13] is an autosomal recessive disorder of epithelial ion transport caused by mutations in the CF transmembrane conductance regulator gene (CFTR), and is characterized by chronic obstructive lung disease, bronchiectasia, and exocrine pancreatic insufficiency.

Candidate genes were chosen from chr7q22.1-31.33 that contains 110 genes including CFTR. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas the top nine genes have a significant p-value ($\alpha$=0.05). CFTR was ranked in the seventh position with a significant p-value (0.046).

In the ranking we obtained seven genes that were significant but not involved in CF or phenotype related diseases. However, out of the top 25 ranked genes we detected four genes that are known to be linked to CF [19-22].

**Supplementary Table S4: Top 25 ranked candidate genes in Becker muscular dystrophy (BDM).**

| Rank | Symbol | Score | 2fold-change | p-value | Linkage to phenotype |
|---|---|---|---|---|---|
| 1 | VCX2 | 0.0771 | 0 | 0.0066 | |
| **2** | **DMD** | **0.0747** | **0.34** | **0.0272** | **BMD [14]** |
| 3 | GK | 0.0735 | 0 | 0.0537 | |
| 4 | PRKX | 0.0734 | 0 | 0.0567 | |
| 5 | PCYT1B | 0.0722 | 0.52 | 0.0978 | |
| 6 | GPR143 | 0.0721 | 0 | 0.1032 | |
| 7 | ACE2 | 0.0719 | 0 | 0.1155 | |
| 8 | ASPGX2 | 0.0718 | 0 | 0.1185 | |
| 9 | XK | 0.0713 | 0 | 0.1432 | |
| 10 | VCX | 0.0713 | 0 | 0.1489 | |
| 11 | PHEX | 0.0709 | 0 | 0.1702 | |
| 12 | REPS2 | 0.0709 | 0.14 | 0.1719 | |
| 13 | ASMT | 0.0708 | 0 | 0.1770 | |
| 14 | MAGEB2 | 0.0708 | 0 | 0.1817 | |
| 15 | S100G | 0.0706 | 0 | 0.1950 | |
| 16 | CD99 | 0.0705 | 0 | 0.2051 | |
| 17 | IL1RAPL1 | 0.0705 | 0.22 | 0.2053 | |
| 18 | SHOX | 0.0703 | 0 | 0.2177 | |
| 19 | FAM47A | 0.0702 | 1.09 | 0.2235 | |
| 20 | GPR64 | 0.0702 | 0 | 0.2236 | |
| 21 | ZFX | 0.0702 | 0.27 | 0.2267 | |
| 22 | STS | 0.0702 | 0 | 0.2279 | |
| 23 | SMPX | 0.0702 | 0 | 0.2280 | |
| 24 | ARHGAP6 | 0.0701 | 0 | 0.2309 | |
| 25 | PIR | 0.0701 | 0 | 0.2323 | |

Becker muscular dystrophy [14] is a X-linked progressive myopathy caused by mutations within the DMD gene, and is characterized by muscle wasting and weakness, and in some cases with mental impairment.

Candidate genes were chosen from chrXp22.33-21.1 that contains 116 genes including DMD. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas the top two genes have significant p-values ($\alpha$=0.05). DMD ranked on the second position with a significant p-value (0.0272). The other significant gene is not involved in BMD or in a phenotype related disease.

**Supplementary Table S5: Top 25 ranked candidate genes from chr5q11.2 in Stein-Leventhal syndrome**.

| Rank | Symbol | Score | 2fold-change | p-value | Linkage to phenotype |
|---|---|---|---|---|---|
| 1 | DDX4 | 0.0524 | 0.54 | 0.0135 | |
| **2** | **FST** | **0.0424** | **0.36** | **0.0560** | **Stein-Leventhal syndrome [25]** |
| 3 | ESM1 | 0.0396 | 0.42 | 0.1176 | |
| 4 | GZMA | 0.0388 | 0.07 | 0.1388 | |
| 5 | RAB3C | 0.0379 | 0 | 0.1939 | |
| 6 | ITGA1 | 0.0373 | 0 | 0.2210 | |
| 7 | ISL1 | 0.0363 | 0.34 | 0.2693 | |
| 8 | ITGA2 | 0.0356 | 0.20 | 0.3133 | |
| 9 | IL6ST | 0.0348 | 0.11 | 0.3528 | |
| 10 | PPAP2A | 0.0338 | 0.07 | 0.4100 | |
| 11 | HSPB3 | 0.0325 | 0.21 | 0.4814 | |
| 12 | PLK2 | 0.0324 | 0.45 | 0.4851 | |
| 13 | GZMK | 0.0317 | 0.14 | 0.5290 | |
| 14 | IL31RA | 0.0312 | 0 | 0.5538 | |
| 15 | PDE4D | 0.0308 | 0.63 | 0.5783 | |
| 16 | GPBP1 | 0.0295 | 0 | 0.6448 | |
| 17 | MAP3K1 | 0.0293 | 0 | 0.6524 | |
| 18 | CCNU | 0.0282 | 0 | 0.7075 | |
| 19 | ARL15 | 0.0264 | 0.45 | 0.7765 | |
| 20 | SNAG1 | 0.0262 | 0 | 0.7884 | |
| 21 | DHX29 | 0.0238 | 0.09 | 0.8556 | |
| 22 | CDC20B | 0.0233 | 0 | 0.8703 | |
| 23 | MOCS2 | 0.0210 | 0.02 | 0.9142 | |
| 24 | NDUFS4 | 0.0195 | 0.01 | 0.9366 | |
| 25 | SKIV2L2 | 0.0169 | 0.04 | 0.9787 | |

Stein-Leventhal syndrome [23] is a oligogenic hormonal disorder among women putatively related with the FST gene, and is characterized by hyperandrogenism, chronic anovulation and associated with obesity.

Candidate genes were chosen from chr5q11.2 that contains 25 genes including the candidate gene FST [25]. These candidate genes were ranked by our new approach, whereas only the top gene has a significant p-value ($\alpha$=0.05). FST was ranked in the second position with a p-value of 0.056, and we received only one significant gene (DEAD box 4) that is a plausible candidate gene for Stein-Leventhal syndrome.

**Supplementary Table S6: Top 25 ranked candidate genes from chr19p13.2 in Stein-Leventhal syndrome.**

| Rank | Symbol | Score | 2fold-change | p-value | Linkage to phenotype |
|---|---|---|---|---|---|
| 1 | OR7G2 | 0.0546 | 0 | 0.0065 | |
| 2 | EDG8 | 0.0529 | 0 | 0.0129 | |
| 3 | ZNF414 | 0.0465 | 0 | 0.0249 | |
| 4 | ADAMTS10 | 0.0442 | 0 | 0.0382 | |
| **5** | **FBN3** | **0.0423** | **0** | **0.0595** | **Stein-Leventhal syndrome [27]** |
| 6 | ZNF653 | 0.0413 | 0 | 0.0754 | |
| 7 | LASS4 | 0.0406 | 0.66 | 0.0894 | |
| 8 | ACP5 | 0.0406 | 0.15 | 0.0910 | |
| 9 | EMR4 | 0.0397 | 0 | 0.1141 | |
| 10 | ZNF177 | 0.0394 | 0.03 | 0.1233 | |
| 11 | MBD3L1 | 0.0387 | 0 | 0.1464 | |
| 12 | MUC16 | 0.0385 | 0.31 | 0.1529 | |
| 13 | EMR1 | 0.0385 | 0.46 | 0.1558 | |
| 14 | ZNF491 | 0.0383 | 0 | 0.1677 | |
| 15 | CCL25 | 0.0377 | 0.54 | 0.2056 | |
| 16 | COL5A3 | 0.0375 | 0.11 | 0.2140 | |
| 17 | ELAVL3 | 0.0373 | 0.27 | 0.2220 | |
| 18 | ZNF559 | 0.0371 | 0 | 0.2319 | |
| 19 | CD320 | 0.037 | 0.50 | 0.2376 | |
| 20 | STXBP2 | 0.0362 | 0.19 | 0.2778 | |
| 21 | C19orf45 | 0.0362 | 0 | 0.2796 | |
| 22 | CLEC4G | 0.0361 | 0 | 0.2838 | |
| 23 | ZNF44 | 0.0361 | 0.26 | 0.2870 | |
| 24 | MCOLN1 | 0.036 | 0.11 | 0.2902 | |
| 25 | TMED1 | 0.0358 | 0.09 | 0.3040 | |

Stein-Leventhal syndrome [23] is a oligogenic hormonal disorder among women putatively associated with the FBN3 gene, and is characterized by hyperandrogenism, chronic anovulation and associated with obesity.

Candidate genes were chosen from chr19p13.2 that contains 100 genes including the candidate gene FBN3 [27]. These candidate genes were ranked by our new approach, and the top 25 ranked candidate genes are presented here, whereas only the top gene has a significant p-value ($\alpha$=0.05). FBN3 was ranked in the fifth position with a p-value of 0.0595.

**Supplementary Table S7: CPU times for computing the Laplacian Exponential Diffusion Kernel by its definition, the CD and the ICD.**

| Number of nodes in the network | CPU time (sec) Def. (Eq. 1) | CPU time (sec) CD (Eq. 2) | CPU time (sec) ICD (Eq. 3) |
|---|---|---|---|
| 1000 | 32 | 25 | 1 |
| 2000 | 254 | 199 | 8 |
| 5000 | 3637 | 3269 | 98 |
| 10000 | 34717 | 26544 | 633 |
| 16566 (full network) | out of memory* | 114953 | 2048 |

CPU times (sec) for computing the Laplacian Exponential Diffusion Kernel by its definition (Equation 1), by CD (Equation 2), and by ICD (Equation 3) with a threshold leading to an error of 7%-10%. The computation were run on a dual Opteron 250 with 16GB RAM.
*To avoid running out of memory we would need a machine with more than 16GB RAM.

**Supplementary Table S8: Rank and Error of resulting Laplacian Exponential Diffusion Kernel computed by its definition, the CD and the ICD.**

| Number of nodes in the network | Rank of matrix Def. (Eq. 1) | Rank of matrix CD (Eq. 2) | Rank of matrix ICD (Eq. 3) | Error ICD |
|---|---|---|---|---|
| 1000 | 999 | 999 | 167 | 10,09% |
| 2000 | 1999 | 1999 | 337 | 8,92% |
| 5000 | 4999 | 4999 | 773 | 7,94% |
| 10000 | 9999 | 9999 | 1328 | 7,27% |
| 16566 (full network) | 16516 | 16516 | 1829 | 7,28% |

Rank and Error of resulting Laplacian Exponential Diffusion Kernel computed by its definition, (Equation 1), by CD (Equation 2), and by ICD (Equation 3) with a threshold leading to an error of 7%-10%. The computation were run on a dual Opteron 250 with 16GB RAM.

**Supplementary Table S9: Example for determining an appropriate neighborhood size using the example of data set 1 (FXS [11])**

| Parameter β | Approx. size of neighborhood | Fisher's Omnibus | $\chi^2_{2k}$ |
|---|---|---|---|
| 0.5 | 20 | 230.79 | 0.113 |
| 0.1 | 80 | 230.72 | 0.113 |
| 0.05 | 150 | **239.90** | **0.064** |
| 0.01 | 500 | 239.34 | 0.067 |
| 0.005 | 1000 | 237.57 | 0.078 |
| 0.001 | 4000 | 231.16 | 0.129 |

The neighborhood size is controlled by a weighting function ( $w = e^{-\beta \cdot r}$ ). Applying the Fisher omnibus meta-analysis ( $S = \sum -2\ln(p - value)$ ) for each parameter β, new p-values are generated from a $\chi^2$ distribution. The parameter β, for which the smallest p-value is observed (here: β = 0.05), leads to the appropriate neighborhood size for FXS (approx. 150 genes).

# Chapter 5

# *Study 4*: Gene prioritization by network analysis of differential expression using machine learning approaches

This study presents algorithmic advancements and optimized algorithms of the method proposed in *Study 3*, again replacing the prior knowledge about the biological process by experimental data on differential gene expression. For doing that, we have proposed three strategies scoring disease candidate genes relying on network-based machine learning (ML) approaches, such as kernel ridge regression, heat kernel, and Arnoldi kernel approximation. For comparison purposes, a local measure based on the expression of the direct neighbors is also computed. For assessing the performance of these algorithms, we have benchmarked these strategies on mouse knockout experiments. Results revealed that all suggested algorithms could successfully rank the sought gene on high ranking positions and outperform a standard procedure in genetics that ranks candidate genes based solely on their differential expression levels.

**Publication**
This work has been published in 2010 as an electronic version in the *BMC Bioinformatics* journal (ISI IF 2009: 3.4).
*Reference:* Nitsch D, Goncalves JP, Ojeda F, de Moor B, Moreau Y (2010). Candidate Gene Prioritization by Network Analysis of Differential Expression using Machine Learning Approaches. *BMC Bioinformatics* 11:460. PMID: 20840752.

**Personal contribution**

The PhD candidate has developed the algorithms in collaboration with the first and second co-author. The PhD candidate has designed the study, implemented the algorithms in Matlab, set up the benchmark data set, preprocessed and normalized the experimental data, analyzed the results, and drafted the paper.

Nitsch et al. BMC Bioinformatics 2010, 11:460
http://www.biomedcentral.com/1471-2105/11/460

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

# Candidate gene prioritization by network analysis of differential expression using machine learning approaches

Daniela Nitsch[1*], Joana P Gonçalves[1,2,3], Fabian Ojeda[1], Bart de Moor[1], Yves Moreau[1]

## Abstract

**Background:** Discovering novel disease genes is still challenging for diseases for which no prior knowledge - such as known disease genes or disease-related pathways - is available. Performing genetic studies frequently results in large lists of candidate genes of which only few can be followed up for further investigation. We have recently developed a computational method for constitutional genetic disorders that identifies the most promising candidate genes by replacing prior knowledge by experimental data of differential gene expression between affected and healthy individuals.
To improve the performance of our prioritization strategy, we have extended our previous work by applying different machine learning approaches that identify promising candidate genes by determining whether a gene is surrounded by highly differentially expressed genes in a functional association or protein-protein interaction network.

**Results:** We have proposed three strategies scoring disease candidate genes relying on network-based machine learning approaches, such as kernel ridge regression, heat kernel, and Arnoldi kernel approximation. For comparison purposes, a local measure based on the expression of the direct neighbors is also computed. We have benchmarked these strategies on 40 publicly available knockout experiments in mice, and performance was assessed against results obtained using a standard procedure in genetics that ranks candidate genes based solely on their differential expression levels (*Simple Expression Ranking*). Our results showed that our four strategies could outperform this standard procedure and that the best results were obtained using the *Heat Kernel Diffusion Ranking* leading to an average ranking position of 8 out of 100 genes, an AUC value of 92.3% and an error reduction of 52.8% relative to the standard procedure approach which ranked the knockout gene on average at position 17 with an AUC value of 83.7%.

**Conclusion:** In this study we could identify promising candidate genes using network based machine learning approaches even if no knowledge is available about the disease or phenotype.

## Background

Discovering novel disease genes is still challenging for constitutional genetic diseases (a disease involving the entire body or having a widespread array of symptoms) for which no prior knowledge - such as known disease genes or disease-related pathways - is available. Performing genetic studies frequently result in large lists of candidate genes of which only few can be followed up for further investigation. *Gene prioritization* establishes the ranking of candidate genes based on their relevance with respect to a biological process of interest, from which the most promising genes can be selected for further analysis.

Several computational methods for prioritizing candidate genes have been proposed, such as ENDEAVOUR [1], GeneWanderer [2], or Prioritizer [3] that rank candidate genes based on associations between known disease genes and candidate genes using different data sources and methodology. *Lage et al. (2007)* developed a phenome-interactome network that integrates phenotypic literature information from OMIM with a cross species protein-protein interaction (PPI) network [4]. *Chen*

* Correspondence: Daniela.Nitsch@esat.kuleuven.be
[1]Department of Electrical Engineering (ESAT-SCD) Katholieke Universiteit Leuven, 3001 Leuven, Belgium
Full list of author information is available at the end of the article

BioMed Central

© 2010 Nitsch et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

*et al. (2009)* applied link based strategies widely used in social and web network analyses to prioritize disease candidate genes based on PPI networks [5]. *Subramanian et al. (2005)* developed a computational method, Gene set enrichment analysis (GSEA), that determines whether an *a priori* defined set of genes shows statistically significant concordant differences between two biological states (e.g. phenotypes), i.e. whether the genes are differentially expressed or not [6]. Similarly, *Liu et al. (2007)* developed a network-based approach, Gene Network Enrichment Analysis (GNEA) for the identification of transcriptionally altered biological processes between disease and normal states [7]. It analyzes gene expression microarray and protein-protein interaction data to identify the affected regions in a protein-protein interaction network. Beside these methods, others can be found in the literature that prioritize candidate genes for human diseases. Among these, several were implemented as web-based applications that can be freely accessed. In a previous study we have reviewed distinct web-based and freely accessible gene prioritization tools for human diseases [8].

The aforementioned methods usually rank candidate genes by matching their information across multiple data sources against a profile derived from a set of genes, keywords, pathways, or biological processes already known to be associated with the phenotype. However, often only little is known about the molecular basis of the phenotype, such as no known disease genes and limited knowledge about the biological cascades involved, and only few or no keywords are known which can be used either to retrieve genes or to perform text mining. To overcome this limitation, we proposed in a previous study a computational method to identify the most promising candidates within a region for which limited or no prior knowledge is available regarding a phenotype of interest by using experimental data on differential gene expression between affected and healthy individuals [9]. Using a network-based approach, we assessed the relevance of a candidate gene by considering the level of differential expression in its neighborhood under the assumption that strong candidates tend to be surrounded by differentially expressed neighbors. For several genetic diseases, however, there is no guarantee that the expression levels of the disease gene itself is affected, rather, genes ''downstream'' of the disease gene are those whose expression will be affected. Therefore, we consider the differential expression data at the network levels instead of (isolated) gene levels. Mapping expression patterns on a network, we then expect to observe a disrupted expression module around the disease gene, while other candidate genes (not causally related to the phenotype) should not be part of such a disrupted expression module. This approach is less biased than assessing relevant candidate genes by performing text mining, finding associations between known disease genes and candidate genes, or detecting disease relevant pathways. We define a notion of a soft neighborhood where each gene is given a contributing weight, which decreases with the distance from the candidate gene in the network. To account for multiple paths between genes, we defined a notion of distance using the Laplacian exponential diffusion kernel. Finally, we scored candidates by aggregating the differential expression of their neighbors weighted as a function of distance.

To further improve prioritization results we have extended our previous work in this study by applying four different strategies to prioritize candidate genes based on network analysis of differential expression using distinct machine learning approaches to determine whether a gene is surrounded by highly differentially expressed genes in a functional association or protein-protein interaction network. Here, we have mainly focused on further performance improvement and determination of an appropriate neighborhood for network propagation of differential expression analysis by considering and benchmarking many variables occurring here in the presented ranking strategies. We have further compared our prioritization results with a standard procedure in genetics that ranks candidate genes based solely on their differential expression levels.

First, we propose an alternative to our previous idea of defining a notion of distance using the Laplacian exponential diffusion kernel. Instead of aggregating the differential expression of neighbors weighted as a function of distance, we have smoothed a candidate gene's differential expression levels through kernel ridge regression. We name this strategy *Kernel Ridge Regression Ranking*.

Second, we have applied the *Heat Kernel Diffusion Ranking*, introduced *by Chung and Yau (1999)* [10], to our problem of disease candidate gene prioritization.

Third, we have carried out network diffusion by applying the Arnoldi algorithm based on a Kyrlov Space method as presented in [11]. We name this strategy *Arnoldi Diffusion Ranking*.

Fourth, we have ranked the candidate genes by combining their differential expression levels with the average of the differential expression levels among their direct neighbors in a functional association or protein-protein interaction network, which we name *Direct Neighborhood Ranking*. This straightforward approach for scoring candidates is presented here for comparison purposes as a naïve strategy for network analysis of differential expression.

We have benchmarked these four strategies on 40 publicly available data sets originated from Affymetrix chips on which mice with (simple) knockout genes were

tested against controls. The raw *cel* files were downloaded from GEO [12]. For each data set we have computed the differential expression levels for each gene in our network based on the expression in the knockout experiment versus the expression in the control (see Methods).

Since we were seeking for a suitable interaction network for our application, we have considered two different types of networks: functional association networks and protein-protein interaction (PPI) networks. As a functional association network, we have used two different STRING releases (version 7.1 [13] and version 8.2 [14]), which include associations and physical interactions coming from heterogeneous databases. As PPI networks, we have used BioGRID [15] and I2 D [16]. By using different networks in our ranking strategies, we evaluated their suitability as an interaction network for our application, and to what extent performance is influenced by the characteristics of the underlying network.

In the field of cancer, some methods for network analysis are available for gene expression towards identification of expression signatures or of dysregulated subnetworks as biomarkers. However, the task we address is that of prioritization of disease causing genes in constitutional genetic disorders using expression data. That problem is different from that of signature identification in cancer because it focuses on the ranking of candidate genes instead of the identification of subnetworks (see for example [17-19]). There is no established method to tackle the specific problem of candidate gene prioritization in constitutional disorders from expression data.

## Results

In this section, we have evaluated the gene prioritization results obtained using each of the presented strategies, focusing on performance improvement. We aim to assess whether machine learning approaches based on random walks can outperform a standard procedure in genetics, the *Simple Expression Ranking* (see for example [20-22]).

The previous described methods require several parameters that we have tuned in order to obtain stable ranking results. Additionally, we combine several preprocessing techniques, expression measures and networks:

*(1) Preprocessing*
Gene expression data was preprocessed using three different techniques: MAS5, RMA and GCRMA (for details see Methods section).

*(2) Expression measure*
We have computed three different expression measures: log2 ratio, the test statistic (derived from CyberT [23], and the significant log2 ratio (p-value derived from CyberT).

*(3) Network*
Four different mouse networks were selected for diffusion: two versions of the functional association network obtained from STRING (versions 7.1 and 8.2), a PPI network obtained from BioGRID (version 2.0.61), and a PPI network obtained from I2 D (version 1.72).
All strategies were applied to these four networks. The Similarity Network for the *Kernel Ridge Regression Ranking* was obtained using the Laplacian Exponential Diffusion kernel ($\alpha$ = 0.5), the Regularized Commute Time kernel ($\alpha$ = 0.90 and $\alpha$ = 0.95), and the Regularized Laplacian Diffusion kernel ($\alpha$ = 1 and $\alpha$ = 2), see equations (1)-(3). For the *Heat Kernel Diffusion Ranking* we have applied discrete approximation of the Heat Kernel rank approach [24], see equation (9). For the *Arnoldi Diffusion Ranking* we have applied Arnoldi approximation [11], see equation (11). Finally, for the *Direct Neighborhood Ranking* we have used the network directly to capture a neighborhood for a candidate gene, see equation (13).

*(4) Methods and parameter setting*
*Kernel Ridge Regression Ranking*, *Heat Kernel Diffusion Ranking*, and *Arnoldi Diffusion Ranking* require a value for the diffusion rate $\alpha$. We have chosen this parameter to be 0.5 for all strategies.
*Kernel Ridge Regression Ranking* requires two parameters, $\lambda$, for which we have chosen five different values$\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$, and *nn*, the maximum number of neighbors, for which we have chosen three values: 30, 50, and 100.
*Heat Kernel Diffusion Ranking* requires one parameter, *m*, defining the number of iterations, i.e., the number of steps in the random walk through the network. *Arnoldi Diffusion Ranking* requires one parameter, *m*, defining the number of iterations for obtaining an approximation of the network diffusion using the Arnoldi algorithm.
*Direct Neighborhood Ranking* requires a weight, *a*, for which we chose a value of 0.5, so that both the expression of the candidate itself and its average surrounding expression in the network were weighted equally. As potential values for $\varepsilon$, we have chosen 0.15, 0.4 and 0.7 (regarding the confidence scores in STRING).

### Validation

The parameter settings were tuned to obtain stable ranking results for the benchmark data, indicated by the rank of the knockout gene. We have computed ROC

curves and the corresponding AUC values and counted the number of knockout genes ranked in the top10% for all ranking lists using all parameter settings. A description of the parameter tuning procedure follows. In this step, we used the *STRING network* (v.7.1) in all the presented strategies. After the determination of the optimal parameters, we have explored alternative kernels in the *Kernel Ridge Regression Ranking*, different preference vectors in the *Heat Kernel Diffusion Ranking*, and distinct networks in all strategies using the determined parameter settings.

#### Parameter Tuning

Table S1a (see additional file 1) illustrates the ranking results obtained by applying the *Kernel Ridge Regression Ranking* using different values for $\lambda$ and *nn* and the Laplacian Exponential Diffusion Kernel with $\alpha$ = 0.5 using the *STRING network* (v.7.1). Based on the results presented we have chosen $\lambda$ to be 1 and *nn* to be 50 since these values lead to the most stable ranking results in comparison to other values, considering the AUC and the number of top ranked knockout genes.

Table S1c (see additional file 1) illustrates the ranking results obtained by applying the *Heat Kernel Diffusion Ranking* using distinct number of steps in the random walk through the *STRING network* (v.7.1). These results clearly show that the method is able to generate a reliable ranking in a reduced number of steps, i.e. $m$ = 2.

Table S1f (see additional file 1) illustrates the ranking results obtained by applying the *Arnoldi Diffusion Ranking* using distinct number of iterations in the Arnoldi approximation and the *STRING network* (v.7.1). The results clearly show that the method is able to produce a reliable ranking using only 2 iterations, as in the case of the *Heat Kernel Diffusion Ranking*, i.e. $m$ = 2.

Table S1e (see additional file 1) illustrates the ranking results obtained by applying the *Direct Neighborhood Ranking* using the *STRING network* (v.7.1) and $a$ = 0.5 with distinct thresholds $\varepsilon$. Results show a decrease in performance with the increase of $\varepsilon$, due to information loss (many associations are then missing from the network), i.e. $\varepsilon$ = 0.15.

Table 1 shows the overall results of the four strategies including their best performing parameters, and using two different releases of the *STRING network* (version 7.1 and version 8.2).

#### Kernels in Kernel Ridge Regression Ranking

The *Kernel Ridge Regression Ranking* is based on a kernel matrix derived from the *STRING network* or the PPI network from BioGRID or I2 D, as shown in equations (1)-(4). As discussed in the Methods section, several kernel matrices are available in the literature. In this study, we have assessed which of the three considered kernels performs best in our application. Table S1b (see additional file 1) shows the results based on the Laplacian Exponential Diffusion Kernel, Regularized Laplacian Kernel and Regularized Commute time Kernel with different parameters. It stands out that the Laplacian Exponential Diffusion Kernel with $\alpha$ = 0.5 performs better than the other kernels. Thus, in our study, the Laplacian Exponential Diffusion kernel with $\alpha$ = 0.5 will be used as the kernel matrix in the *Kernel Ridge Regression Ranking*.

#### Preference vector initialization in Heat Kernel Diffusion Ranking

*Francisco et al. (2009)* suggested initializing the preference vector $p_0$ with binary values [25]: the seed genes known to be involved in the disease would be set to *1*, as all other genes in the vector would be set to *0*. In order to assess the contribution of the gene expression levels to the ranking, we have compared four different scenarios: first, only the candidate genes were initialized with *1* and the remaining genes with *0* (same procedure as in [25]), independent of their expression value. Second, we have only initialized the candidate genes with their differential expression levels obtained in the experiment and the remaining genes with *0*. Third, we have filled the preference vector $p_0$ with all the expression values available in the experiment. Fourth, we have initialized all genes that are differentially expressed in the experiment with *1*, all other genes with *0*, thus we have made no difference between genes that were highly or weakly differentially expressed. Table S1d (see additional file 1) shows the result of these four scenarios for initializing the preference vector.

It can be seen that by using only binary values (scenario 1), the *Heat Kernel Diffusion Ranking* performs poorly (AUC = 61.2% with only 5 knockout genes in the top 10%). However, using expression values instead of binary ones, the method performs better. This shows that the contribution of expression levels coming from a disease or knockout related microarray experiment is significant. Furthermore, if all available expression values are added to the preference vector (scenario 3), the results were further improved against the ones obtained by initializing the preference vector with the expression levels of the candidate genes only (scenario 2). However, if the expression values are replaced by binary ones (scenario 4), the results are slightly worse than considering differential expression values. By considering all available expression values (scenario 3), the ranking obtained between 31 and 32 top 10% ranked knockout genes with an AUC of between 88.9% and 90.7% for the log2 ratio or the test statistic as the expression measure, independent of the preprocessing technique. In subsequent steps, we use all available expression data for initializing

**Table 1 Overview of prioritization results**

| | | | STRING v. 7.1 | | | STRING v. 8.2 | | |
|---|---|---|---|---|---|---|---|---|
| | | | **top 10** | **top 20** | **AUC** | **top 10** | **top 20** | **AUC** |
| | **Standard genetic procedure: Simple Expression Ranking** | | **20** | **25** | **0.801** | **20** | **25** | **0.801** |
| | **Direct Neighborhood Ranking** | log2 ratio | 27 | 31 | 0.859 | 12 | 23 | 0.747 |
| | $\varepsilon$ >0.15, $a$ = 0.5 | sign. log2 ratio | 28 | 31 | 0.880 | 12 | 24 | 0.760 |
| | | test statistic | 29 | 30 | 0.856 | 13 | 22 | 0.738 |
| | **Kernel Ridge Regression Ranking** | log2 ratio | 23 | 29 | 0.809 | 14 | 26 | 0.759 |
| | $\lambda$ = 1, $nn$ = 50, $K_{LED}$ , $^{\alpha = 0.5}$ | sign. log2 ratio | 27 | 32 | 0.868 | 17 | 25 | 0.817 |
| **RRMA** | test statistic | 20 | 26 | 0.771 | 15 | 20 | 0.691 | |
| | **Heat Kernel Diffusion Ranking** | log2 ratio | **32** | 34 | 0.900 | **32** | 33 | 0.913 |
| | all expression values for $p_0$ , $m$ = 2, $\alpha$ = 0.5 | sign. log2 ratio | 31 | 34 | **0.910** | 31 | 35 | **0.923** |
| | | test statistic | **32** | 34 | 0.901 | **32** | 34 | 0.911 |
| | **Arnoldi Diffusion Ranking** | log2 ratio | 27 | 31 | 0.857 | 27 | 29 | 0.851 |
| | $m$ = 2, $\alpha$ = 0.5 | sign. log2 ratio | 28 | 31 | 0.885 | 28 | 31 | 0.873 |
| | | test statistic | 28 | 30 | 0.855 | 28 | 29 | 0.844 |
| | **Standard genetic procedure: Simple Expression Ranking** | | **18** | **24** | **0.777** | **18** | **24** | **0.777** |
| | **Direct Neighborhood Ranking** | log2 ratio | 27 | 31 | 0.874 | 12 | 23 | 0.761 |
| | $\varepsilon$ >0.15, $a$ = 0.5 | sign. log2 ratio | 25 | 28 | 0.855 | 11 | 22 | 0.736 |
| | | test statistic | 27 | 31 | 0.863 | 12 | 24 | 0.750 |
| | **Kernel Ridge Regression Ranking** | log2 ratio | 21 | 28 | 0.769 | 17 | 24 | 0.756 |
| | $\lambda$ = 1, $nn$ = 50, $K_{LED}$ , $^{\alpha = 0.5}$ | sign. log2 ratio | 27 | 31 | 0.835 | 20 | 26 | 0.796 |
| **GCRMA** | | test statistic | 19 | 22 | 0.745 | 16 | 23 | 0.744 |
| | **Heat Kernel Diffusion Ranking** | log2 ratio | 31 | 33 | **0.905** | **32** | 34 | **0.914** |
| | all expression values for $p_0$ , $m$ = 2, $\alpha$ = 0.5 | sign. log2 ratio | 28 | 33 | 0.889 | 29 | 34 | 0.907 |
| | | test statistic | **32** | 33 | 0.895 | **32** | 35 | 0.913 |
| | **Arnoldi Diffusion Ranking** | log2 ratio | 27 | 32 | 0.875 | 26 | 31 | 0.874 |
| | $m$ = 2, $\alpha$ = 0.5 | sign. log2 ratio | 25 | 29 | 0.860 | 25 | 28 | 0.852 |
| | | test statistic | 27 | 31 | 0.865 | 28 | 31 | 0.862 |
| | **Standard genetic procedure: Simple Expression Ranking** | | **24** | **28** | **0.837** | **24** | **28** | **0.837** |
| | **Direct Neighborhood Ranking** | log2 ratio | 23 | 27 | 0.846 | 9 | 20 | 0.743 |
| | $\varepsilon$ > 0.15, $a$ = 0.5 | sign. log2 ratio | 25 | 28 | 0.844 | 11 | 22 | 0.729 |
| | | test statistic | 27 | 30 | 0.855 | 12 | 24 | 0.745 |
| | **Kernel Ridge Regression Ranking** | log2 ratio | 18 | 24 | 0.736 | 17 | 24 | 0.766 |
| | $\lambda$ = 1, $nn$ = 50, $K_{LED}$ , $^{\alpha = 0.5}$ | sign. log2 ratio | 23 | 29 | 0.834 | 17 | 22 | 0.755 |
| **MAS5** | | test statistic | 13 | 18 | 0.790 | 16 | 24 | 0.790 |
| | **Heat Kernel Diffusion Ranking** | log2 ratio | 26 | 32 | 0.877 | 28 | 34 | 0.890 |
| | all expression values for $p_0$ , $m$ = 2, $\alpha$ = 0.5 | sign. log2 ratio | 26 | 31 | 0.877 | 26 | 34 | 0.899 |
| | | test statistic | **32** | 32 | **0.890** | **31** | 34 | **0.904** |
| | **Arnoldi Diffusion Ranking** | log2 ratio | 25 | 27 | 0.849 | 24 | 27 | 0.851 |
| | $m$ = 2, $\alpha$ = 0.5 | sign. log2 ratio | 25 | 29 | 0.853 | 25 | 27 | 0.847 |
| | | test statistic | 26 | 30 | 0.858 | 27 | 29 | 0.850 |

The results are based on optimized parameter settings for all presented strategies using different *STRING networks*. Note that the results of the *Simple Expression Ranking* do not depend on the network because they are only using differential expression levels to rank the candidate genes.

the preference vector in the *Heat Kernel Diffusion Ranking*.

### Error Reduction

Table 1 shows that the *Simple Expression Ranking* performs well: using MAS5 preprocessed data we have obtained an AUC of 83.7%. However, by applying the presented machine learning strategies based on random walks, the results could be further improved.

Results of the *Heat Kernel Diffusion Ranking* using RMA preprocessed data have obtained an AUC value of 91% and 92.3% using the *STRING networks* v7.1 and v8.2, respectively, and the significant log2 ratio as expression measure. This presents an error reduction of

52.8% relative to the *Simple Expression Ranking* for the *STRING network* v8.2 and 44.8% for the *STRING network* v7.1. Regarding the number of knockout genes, the *Heat Kernel Diffusion Ranking* has ranked a maximum of 32 knockout genes in the top 10% for both *STRING network* releases using the test statistic as expression measure, independent of the preprocessing technique. The *Simple Expression Ranking* could rank at most 24 knockout genes in the top 10% for MAS5 preprocessed data, which again presents an error reduction of 50% using the *Heat Kernel Diffusion Ranking*.

The *Kernel Ridge Regression Ranking* strategy could outperform the *Simple Expression Ranking* only for RMA using the significant log2 ratio as expression measure and *STRING network* v7.1. The corresponding error reduction was of 19% regarding the AUC (86.8%) and the number of knockout genes (27) relative to the *Simple Expression Ranking* using MAS5. For *STRING* v8.2 and all other settings the *Kernel Ridge Regression Ranking* could not outperform the *Simple Expression Ranking*.

The *Arnoldi Diffusion Ranking* strategy could outperform the *Simple Expression Ranking* for RMA preprocessed data using the significant log2 ratio or test statistic as expression measure and *STRING* v7.1 or for GCRMA preprocessed data using the log2 ratio as expression measure and *STRING* v8.2. The error could be reduced by 29.4% in terms of the AUC for *STRING* v7.1 (88.5%) and by 22.7% for *STRING* v8.2 (87.4%). Regarding the number of top ranked knockout genes, we could reduce the error for both networks by 25% (28 top ranked knockout genes) relative to the *Simple Expression Ranking*. For some other settings this strategy could outperform the *Simple Expression Ranking* as well.

The *Direct Neighborhood Ranking* strategy could outperform the *Simple Expression Ranking* for RMA preprocessed data using the significant log2 ratio or test statistic as expression measure and the *STRING network* v7.1. The error could be reduced by 26.4% in terms of the AUC (88%) and by 31% in terms of the number of knockout genes (29) relative to the *Simple Expression Ranking* using MAS5 preprocessed data. For some other settings this strategy could outperform the *Simple Expression Ranking* as well. However, for *STRING* v8.2 the *Direct Neighborhood Ranking* could not outperform the *Simple Expression Ranking*.

As a consequence, the *Heat Kernel Diffusion Ranking* shows the largest error reduction relative to the *Simple Expression Ranking* considering both the AUC values and the number of top ranked knockout genes. Furthermore, these results were achieved independent of the preprocessing technique and for both *STRING network* releases.

#### Analysis of dependency on expression levels

We have further analyzed the influence of the expression levels of the candidate genes in the performance of the four presented ranking strategies. For this purpose, we have set the expression levels of the knockout genes to *0* and compared with the original results. Table S2 shows that none of the strategies achieves a comparable level of performance if the expression level of the knockout gene is set to *0*. Therefore it is reasonable to assume that the performance of all four strategies in this study is highly dependent on integration of the expression levels of the knockout gene.

#### Comparison of different networks: functional association networks vs. protein-protein interaction networks

As shown in the Methods Section, the mouse PPI network originated from BioGRID contains substantially less information then both mouse *STRING networks*. Nevertheless, we have analyzed to what extent the gene prioritization methods depend on the choice of the network and how sparsity is influencing the results.

Table 1 illustrates a direct comparison between both STRING releases using the four gene prioritization strategies. It shows that the performance of both, the *Kernel Ridge Regression Ranking* and *Direct Neighborhood Ranking*, using *STRING* v7.1 is better than using *STRING* v8.2. By updating the *STRING network* from version 7.1 to 8.2, the error was increased by approx. 30%-50% in comparison to version 7.1, depending on the parameter setting. However, the performance for the *Heat Kernel Diffusion Ranking* and *Arnoldi Diffusion Ranking* applying both STRING releases are comparable, even though the releases differ substantially from each other. This demonstrates that these two strategies are not as dependent on the choice of the network as the other two. On the other hand, by updating the *STRING network* from version 7.1 to 8.2 the performance could not be improved further using the *Heat Kernel Diffusion Ranking* or *Arnoldi Diffusion Ranking* as expected.

Figure 1 shows the ROC curves of the performance of the presented strategies for gene prioritization using RMA preprocessed data and the significant log2 ratio as the expression measure for both STRING network releases in comparison to the *Simple Expression Ranking* using MAS5 preprocessed data. The reason for choosing RMA as the preprocessing technique and the significant log2 ratio as the expression measure is that this setting leads generally to the most stable and reliable result in our study (see Table 1).

Using the mouse PPI network obtained from the BioGRID database, we could not obtain good quality ranking results for our benchmark. Out of the 40 data sets, only six could be used on the PPI network because the knockout genes of the other 34 data sets were absent

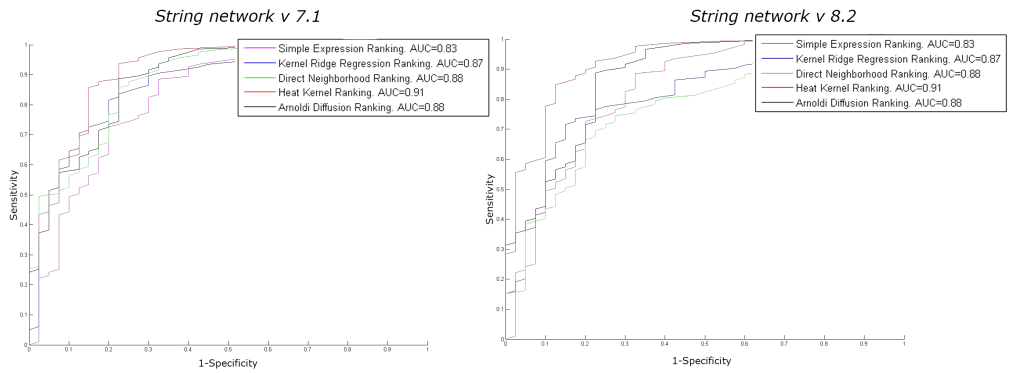



**Figure 1 Performance comparison of *STRING network* v 7.1 and v 8.2**. Comparison of the performance between *STRING network* version 7.1 and version 8.2 applying the four presented strategies using RMA preprocessed data and the significant log2 ratio as the expression measure in comparison to the *Simple Expression Ranking* using MAS5 preprocessed data.

from the PPI network (this network is very sparse, see Table 2). For validating the performance of the strategies, we have used 100 candidate genes that were ranked for each data set. However, using this PPI network, only 4-10 of the candidate genes were found in the network (data not shown). Therefore, we could not perform a meaningful comparison between the functional association network from STRING and the PPI network from BioGRID. Thus, we can draw the conclusion that a sparse network, like the BioGRID mouse network, is not suitable for our application.

The mouse PPI network from the I2 D database is a higher-coverage PPI network (see Table 2). However, like for the BioGRID mouse network, the I2 D mouse network is missing knockout genes from the benchmark. Only 35 out of 40 data sets could be used on the PPI network because the others were incomplete (data not shown). For this reason we have extended the network by five additional nodes which are isolated from the other nodes in the networks, since the database does not provide information about interactions between these nodes and the others in the network. Table S3 (see additional file 1) shows that by using the I2 D network the results coming from the *Heat Kernel Diffusion Ranking* and the *Arnoldi Diffusion Ranking* were only slightly better than those obtained by the *Simple Expression Ranking*, for the other ranking strategies the I2 D network performed worse in terms of the AUC value and top ranked knockout genes. Comparing with the *STRING network*, our results have shown that a functional association network like the *STRING network* performs better in our application than a PPI network.

### Comparison of different expression measures and preprocessing techniques

In this study, we have evaluated not only different networks applied on the four presented strategies, but also different expression measures in combination with the preprocessing techniques applied by the strategies. Table 1 illustrates that in our application RMA best results correlates with the significant log2 ratio as expression measure, GCRMA with the log2 ratio, and MAS5 with test statics. In terms of the presented prioritization strategies, Table 1 shows that the *Kernel Ridge Regression* best results correlate with the significant log2 ratio as expression measure, and the *Heat Kernel Diffusion Ranking* with test statistic and significant log2 ratio.

These results show that there is no clear conclusion to draw in terms of identifying the best performing preprocessing technique or expression measure. Finding the best performing preprocessing technique is a challenge and has been widely discussed in the literature, such as [26-29].

**Table 2 Overview of the global network properties of the underlying networks**

| Database (mouse) | Number of Genes | Number of Interactions | Average Node Degree |
|---|---|---|---|
| STRING v7.1 | 16,566 | 820,177 | 49.5 |
| STRING v8.2 | 24,442 | 1,405,375 | 57.5 |
| BioGRID v2.0.61 | 1,417 | 2,026 | 2.5 |
| I2 D v1.72 | 10,867 | 79,088 | 10.6 |

## Discussion

As we have introduced in this study, there are several methods in the literature to prioritize candidate genes. Among these approaches, there are methods requiring knowledge about disease-gene association, whereas others do not have this precondition. However, these methods will be ineffective if no knowledge is available for a specific disease. In our previous study, we introduced an approach to overcome this limitation by replacing this knowledge by disease-specific experimental data [9].

The purpose of this study was to further improve the performance of the prioritization results and the determination of an appropriate neighborhood for network analysis which has been a major issue in our previous work. We have extended our previous work extensively by suggesting four strategies to prioritize candidate genes based on distinct machine learning approaches to determine whether a gene is surrounded by highly differentially expressed genes in a network, by considering variations in parameter settings, usage of different kernel functions, preprocessing techniques and expression measurements. All prioritization results coming from these different settings were benchmarked on 40 mouse knockout data sets computing AUC values and the number of knockout genes ranked in the top 10%. The performance was assessed against results obtained using a standard procedure in genetics that ranks candidate genes based solely on their differential expression levels (*Simple Expression Ranking*) that was clearly outperformed by the here presented ranking strategies in terms of AUC values and ranking positions of knockout genes (see Results section).

We have used three distinct random walk based strategies plus one naïve strategy without network diffusion but a direct neighborhood analysis. The random walk strategies all base on the Exponential Diffusion Kernel, although the methodology is distinct: the *Kernel Ridge Regression Ranking* computes the (convergent) kernel to solve a regression problem, the *Heat Kernel Diffusion Ranking* computes an iterative diffusion with only 2 steps over the network and combines it with expression data coming from the experiment which is the Heat Kernel rank, and the *Arnoldi Diffusion Ranking* combines an approximation of the kernel with expression data. It has to be noted that the convergence in the *Heat Kernel Diffusion Ranking* and in the *Arnoldi Diffusion Ranking* has been measured in terms of ranking results and not in terms of kernel assimilation which makes it not comparable to the kernel matrix computation resulting from the *Kernel Ridge Regression Ranking*.

Although we have considered and benchmarked many variables occurring in our ranking strategies, there are still a few that influence the performance of our prioritization strategies.

First, the quality and coverage of the network around the actual disease gene will be a strict bottleneck as we already analyzed in [9]. For example, an isolated gene with no edges in the network can not be effectively prioritized by our methods. We have used the *STRING network* based on the assumption that a network, integrating multiple heterogeneous data sources, is more complete and robust and therefore less prone to the problems caused by sparsity. Further, we must note that extending the data sources from version 7 to 8 has resulted in a larger network with good coverage. Errors in the network might cause both false negatives and false positives.

Continuing improvements in the quality of protein association networks are expected to contribute to increased effectiveness of the proposed method which could be demonstrated by comparing the two releases STRING v7.1 and STRING v8.2. In the latter release the performance could be increased further from 91% to 92.3% using the *Heat Kernel Diffusion Ranking*, although the performances obtained by using both releases are not as significant as by using the *Kernel Ridge Regression Ranking* and *Direct Neighborhood Ranking*. The performance for the *Kernel Ridge Regression Ranking* and *Direct Neighborhood Ranking* using the *STRING network* v7.1 was better than using the *STRING network* v8.2. By updating the *STRING network* from version 7.1 to 8.2, the error was increased by approx. 30%-50%, dependent of the setting.

Worsening in the performance using the updated version of the *STRING network* in the *Direct Neighborhood Ranking* could be explained by the fact that we could observe more neighbors for the candidate genes so that a few highly differentially expressed genes lost their influence in comparison to version 7.1 in which fewer neighbors were detected. For the *Kernel Ridge Regression Ranking*, however, the similarity matrices of both STRING releases differ substantially from each other and cannot be compared directly since there is only a small overlap between both *STRING networks*.

To evaluate to what extent the coverage of the interaction network influences our method, we have further applied 2 distinct PPI networks of different coverage. Using the BioGRID network, results have shown that this network was too sparse for our application because most of the knockout genes from the data sets and candidate genes to be ranked in the underlying study were absent from the network (data not shown). Using the I2 D network, results have shown that we could slightly outperform our Baseline, the *Simple Expression Ranking*. Nevertheless, the *STRING network* is more densely

connected and contains more genes (see Table 2), which led to a better performance for the ranking results.

Like in the BioGRID network, some knockout genes were also absent from the I2 D network. Therefore we extended the network by these nodes, but with no interaction between them and the other nodes in the network. Our results have shown that some of the isolated genes could be ranked highly (data not shown) despite their missing interactions because the ranking strategies also consider the candidate genes' expression levels. As long as an isolated gene has a large differential expression level, it can be ranked highly. However, this is not a guarantee to detect this gene among the top ranked genes because a highly differentially expressed neighborhood will still dominate. Therefore an isolated gene cannot be reliably prioritized by our ranking strategies.

Nevertheless, the ranking strategies using the *STRING network* have performed better than using the I2 D network, probably because of the difference in coverage. One may argue that it is not appropriate to compare a PPI network with a functional association network because the latter is including associations apart from physical interactions, but we are seeking for a suitable and high-coverage interaction network for our application that leads to reliable ranking results of candidate genes, regardless of the nature of the associations.

In the literature, several databases can be found to construct an interaction network that can be categorized in terms of types and scope of data sources, types of interactions, or range of details on protein interactions [30]. Comprehensive protein interaction databases, such as BioGRID [15], DIP [31], BIND [32], IntAct [33], MINT [34], I2 D [16], iRefIndex [35] or STRING [13,14] collect physical interactions or functional associations between proteins, integrating distinct information sources about protein interactions coming from high-throughput experimental data, structural data, manual curation, or functional predictions. Specialized protein interaction databases, such as MIPS [36] or HPRD [37] collect (manually) curated interactions from yeast or human, respectively. For this study, the BioGRID database was used to represent a (sparse) PPI network containing only high-throughput experimental data, the I2 D database was used to represent a high-coverage PPI network integrating different information sources, and STRING was used to represent a high-coverage functional association network combining associations and physical interactions coming from distinct data sources.

In *Nitsch et al. (2009)* we adopted the hypothesis from *Köhler et al. (2008)* that global network-similarity measures capture relationships between disease proteins better than algorithms based on direct interactions [2,9]. The latter was considered in the *Direct Neighborhood Ranking* strategy which performed worse than the *Heat Kernel Diffusion Ranking* leading to a global diffusion measure. Since we have proposed the *Direct Neighborhood Ranking* as a naïve method for comparison purposes, we have already expected not to outperform the other three random walk based strategies because of its systematic bias favoring highly connected genes in the network (which we discuss later).

Second, the integration of expression data proved to have a significant influence on the performance of the prioritization strategies which is supported by the results using the mouse knockout experiments data. By using mouse knockout data sets, we could guarantee to use disease relevant tissue which is an important issue in identifying disease related genes (disease genes might not be expressed in other tissues). We have further analyzed the dependency of the ranking performance on the preprocessing of expression data. In the literature there are many controversial discussions about the performance of these, including MAS5, RMA, and GCRMA, and the community cannot harmonize (see for example [26-29]). For this reason we have applied MAS5, RMA, and GCRMA on the benchmark data sets leading to different ranking results, as we would have expected.

In the Results section we have claimed that we could not draw a clear conclusion in terms of identifying the best performing preprocessing technique or expression measure. Nevertheless, the best results were achieved using RMA preprocessed data in comparison to the other preprocessing techniques. However, in other applications alternative techniques might perform better. The same matters for finding an appropriate expression measure. In the literature there are many measures, from which we have chosen the log2 ratio, the significant log2 ratio, and test statistic derived from CyberT. By analyzing their performance in our application, we could not claim one measure to be outperforming another, however, the significant log2 ratio led to the most stable results over the majority of the settings.

The quality of the experimental data has also an important impact on the performance of the prioritization strategies. By using expression data that was either incomplete or of poor quality, the prioritization strategies could not lead to reliable results. The sensitivity of the different scoring strategies or the underlying networks to this effect may vary (for example, PPI networks have been reported to be more sensitive to this effect). This issue may come up when applying the presented method to real genetic disorders in human. In our previous study, we could already detect four known disease genes in human using public available expression data of good quality [9]. Therefore, the bottleneck of our method is again the sparseness and quality of experimental data.

Third, the choice of appropriate parameters for the presented strategies has a large influence on the performance. Therefore we had several parameters and settings to optimize. By using the optimal parameters and settings, we have obtained stable and robust results for the prioritization problem. All presented strategies have outperformed our Baseline, the *Simple Expression Ranking* - as we have expected - for certain settings. The *Heat Kernel Diffusion Ranking* has shown the largest error reduction relative to the *Simple Expression Ranking* in terms of AUC values and number of top ranked knockout genes, independent of the preprocessing technique and for both *STRING network* releases. Our results show that the *Simple Expression Ranking* is performing well for MAS5 preprocessed data. However, the *Heat Kernel Diffusion Ranking* on RMA preprocessed data could achieve and error reduction of 52.8% for the *STRING network* v8.2 and 44.8% for the *STRING network* v7.1.

Fourth, the fact that a disease related gene may have only a weakly differentially expressed neighborhood negatively affects the prioritization methods. In this case, our hypothesis of observing a strongly differentially expressed neighborhood for disease related genes can fail. By introducing a better experimental design that can trigger a disease pathway more reliably in which the effect is more focused around the disease gene, would overcome this limitation because in our approach we do not consider pathways but only neighbors that are surrounding a gene in the network. In this case the *Simple Expression Ranking* can perform better as long as the disease gene is highly differentially expressed since this naïve method does not depend on expressed pathways or neighborhoods.

Fifth, a systematic bias using a biological network favoring highly connected genes can be observed in our method which leads to sensitivity to skewed degree distributions. This bias is clearly a limitation of our method and can be addressed through proper randomization. By randomizing signals across the network, the bias caused by the higher connectivity (or higher total weights) can be removed since highly connected nodes will have a higher baseline signal across the randomizations. We are currently investigating this issue.

## Conclusion

In this study, we have extended extensively our previous work by applying machine learning approaches based on random walk models to determine whether a gene's neighborhood is highly differentially expressed. We have explored three different random walk based strategies plus one naïve strategy based on a direct neighborhood analysis. These four network-based prioritization strategies for scoring candidate genes based on their differentially expressed neighborhood were benchmarked on 40 publicly available knockout experiments in mice. Performance was assessed against results obtained using a standard procedure in genetics that ranks candidate genes based solely on their differential expression levels (*Simple Expression Ranking*). Results showed that our strategies could outperform this standard procedure and that the best results were obtained using the *Heat Kernel Diffusion Ranking* leading to an average ranking position of 8 out of 100 genes, an AUC value of 92.3% and an error reduction of 52.8% relative to the standard procedure approach which ranked the knockout gene in average at position 17 with an AUC value of 83.7%. Thus, we could identify promising candidate genes using network-based machine learning approaches even if no knowledge is available about the disease or phenotype.

## Methods

### Benchmark data

The benchmark for this study consists of 40 publicly available data sets originated from Affymetrix chips on which mice with knockout genes were tested against controls. The raw *cel* files were downloaded from GEO [12]. Table 3 shows all data sets used in our benchmark.

### Preprocessing

We have preprocessed the gene expression data using three different preprocessing techniques: MAS5 (Affymetrix Microarray Suite 5.0) [38], RMA [39], and GCRMA [40]. Moreover, we have assessed their individual contribution to results seeking the best performance in a number of validation tests.

We have implemented the preprocessing techniques in R using the BioConductor package which includes a large number of meta-data packages available that are oriented towards different types of microarrays. *Affy* is the most common library that is used for data processing and visualization of Affymetrix GeneChip measurements [41].

### Differential expression measures

After the preprocessing step we have computed the differential expression level for each gene in the network based on the expression in the knockout experiment versus the expression in the control for each data set.

We have used three different measures of differential expression:

- log2 ratio that is defined as $\log_2\left(\frac{\text{mutant}}{\text{control}}\right)$.
- Significant log2 ratio: only significant log2 ratios (pvalue < 0.05) are considered, otherwise they are set to 0. The pvalue was computed by CyberT.
- Test statistic: computed from CyberT.

**Table 3 The benchmark data**

| | Gene Name | GEO accession number | | Gene Name | GEO accession number |
|---|---|---|---|---|---|
| 1 | Abca1 | GSE5496 | 21 | Mbnl1 | GSE14691 |
| 2 | Btk | GSE2826 | 22 | Mst1r, Ron | GSE16629 |
| 3 | Cav1 | GSE10849 | 23 | MyD88 | GSE6688 |
| 4 | Cav3 | GSE10848 | 24 | Nos3, eNos | GSE1988 |
| 5 | Cftr | GSE5715 | 25 | Phgdh | GSE8555 |
| 6 | Clcn1 | GSE14691 | 26 | Pmp22 | GSE1947 |
| 7 | Cnr1 | GSE7694 | 27 | PPAR$\alpha$ | GSE6864 |
| 8 | Emd | GSE5304 | 28 | Prkag3, AMPK G3 | GSE4065 |
| 9 | Epas1, Hif-2 | GSE16067 | 29 | Pthlh, Pthrp | GSE17654 |
| 10 | Esrra | GSE7196 | 30 | Rab3a | GSE6527 |
| 11 | Gap43 | GSE12687 | 31 | RasGrf1 | GSE8425 |
| 12 | Gnmt | GSE9809 | 32 | Rbm15 | GSE12628 |
| 13 | Hdac1 | GSE5583 | 33 | Runx | GSE4911 |
| 14 | Hdac2 | GSE6770 | 34 | Scd1 | GSE2926 |
| 15 | Hsf4 | GSE12415 | 35 | Slc26a4 | GSE10587 |
| 16 | Hspa1A, Hsp70.1 | GSE11120 | 36 | Srf | GSE13333 |
| 17 | Il6 | GSE411 | 37 | Tgm2 | GSE10285 |
| 18 | Lhx1, Lim1 | GSE4230 | 38 | Zc3h12a | GSE14891 |
| 19 | Lhx8 | GSE11897 | 39 | Zfp36, Tpp | GSE5324 |
| 20 | Lmna | GSE5304 | 40 | Zfx | GSE7069 |

The benchmark consists of 40 publicly available data sets originated from Affymetrix chips on which mice with (simple) knockout genes were tested against controls.

CyberT [23] employs statistical analyses based on simple t-tests that use the observed variance of replicate gene measurements across replicate experiments, or regularized t-tests that use a Bayesian estimate of the variance among gene measurements within an experiment.

#### Network

A functional protein association network is an undirected graph with proteins as nodes and associations as edges. Whenever there is a functional association between two proteins, an edge will be set between the corresponding nodes in the graph. The weights of the edges represent a confidence value on the evidence of such an association. The functional protein association network for mouse was obtained from STRING [13], a database of known and predicted protein-protein associations derived from heterogeneous data sources and different organisms including both physical interactions and functional associations. In the following we will refer to this network as the *STRING network*.

STRING 7 integrates known interactions coming from interaction databases, such as MINT, HPRD, BioGRID, DIP, Reactome, BIND, and KEGG, as well as text retrieved from PubMed abstracts and other scientific resources [13]. The STRING network v7.1 for the mouse organism contains 16,566 genes and 820,177 interactions with an average node degree of 49.5. STRING 8 [14] includes beside the included databases from the previous release updated databases and new resources, such as IntAct, NCI-Nature Pathway Interaction Database and Gene Ontology protein complexes. The STRING network v8.2 for the mouse organism contains 24,442 genes and 1,405,375 interactions with an average node degree of 57.5. It has to be noted that the two STRING versions share only 375,682 interactions representing 46% and 27% of the v7.1 and v8.2 network sizes respectively, which leads to the fact that these two releases are very different and therefore not comparable. However, we have to mention that these numbers are estimates due to mapping of gene ids.

In addition to two releases of the *STRING network* (versions 7.1 and 8.2), we have also obtained a protein-protein interaction (PPI) network for mouse from BioGRID (version 2.0.61), a repository for physical and genetic interactions derived from literature and high-throughput experiments [15]. The BioGRID v2.0.61 for the mouse organism contains 1,417 genes and 2,026 interactions with an average node degree of 2.5.

We have further used a second PPI network for mouse that was derived from I2 D (Interologous Interaction Database) which is a database integrating existing curation, such as IntAct, BIND, DIP, MINT, and HPRD, high-throughput and predicted interactions [16]. The I2 D v1.72 for the mouse organism contains 10,867 genes and 79,088 interactions with an average node degree of 10.6.

Table 2 shows an overview about the global network properties of underlying databases.

In this study, we have applied network-based strategies to prioritize candidate genes. Further, we have compared the performance of the ranking strategies on both functional association (*STRING network)* and PPI networks using the BioGRID network and the I2 D network as representatives.

In *Nitsch et al. 2009* we hypothesized that global network-similarity measures capture relationships between disease proteins better than algorithms based on direct interactions [9]. We used graph kernels to capture global relationships within a graph, computing global similarity of two nodes as the probability of reaching one node at some time point after a random walk starting from another node. The resulting graph led to a global similarity network where an edge between two nodes did not represent a direct interaction, but rather their similarity in this network. From this similarity network, the distance network could be easily derived.

In the *Kernel Ridge Regression Ranking*, we have further explored the idea of using graph kernels to detect global similarities between any genes in the network by implementing different kernels and comparing their performance. Furthermore, instead of aggregating the differential expression of neighbors weighted as a function of distance, we have smoothed a candidate gene's differential expression level using kernel ridge regression, considering its most similar neighbors and their differential expression.

### Candidate genes

For each data set we have selected a set of 100 candidate genes, including the knockout gene. For getting the candidates, we have chosen the knockout genes' nearest 100 genes on the chromosome based on the BioMart data mining tool which accesses data from the Ensembl Genome Browser.

We have then prioritized this set of candidate genes using four different strategies. Results were evaluated by retrieving the position of the knockout gene in the ranking list and by calculating the corresponding AUC value.

### Evaluation

Evaluation of the different strategies was accomplished by ranking the candidate genes in each data set using a number of different parameter and measure settings: MAS5/RMA/GCRMA combined with log2 ratio/significant log2 ratio/test statistic as expression measure using different kernels in the *Kernel Ridge Regression Ranking* strategy, using different preference vector initializations, iteration numbers and diffusion parameters in the *Heat Kernel Diffusion Ranking*, using different thresholds for the edge weights in the *Direct Neighborhood Ranking*, and different initial vectors and iteration numbers in the *Arnoldi Diffusion Ranking*.

The performance was assessed based on the position of the knockout gene in the ranking list. Ideally, the knockout gene should appear in the top of the ranking list based on the hypothesis that this gene is causing all the disruption in the expression of the genes in the network.

As evaluation measure, we have computed AUC values and the number of knockout genes that were successfully ranked in the top 10% among the candidate genes.

The AUC (Area Under the ROC Curve) is a standard measure of the performance of the ranking algorithm and assesses its ability to separate the two classes "positive ranked" (genes that are highly ranked) and "negative ranked" (genes that are lowly ranked). The corresponding ROC (Receiver Operating Characteristic) curve is achieved by imposing and varying a threshold (from rank 1 to n) to separate these two classes, which leads to *true positive rates* and *false positive rates* for each threshold $k$. To obtain the ROC curve, the FPRs (false positive rate) are plotted against the TPRs (true positive rate). Following *Fawcett (2006*) the AUC is equivalent to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [42]. This is equivalent to the Wilcoxon test of ranks [43]. For instance, an AUC value of 100% indicates that every knockout gene ranks in the first position, whereas an AUC value of 50% means that the knockout genes rank randomly.

### Ranking Strategies

#### *Simple Expression Ranking as a Baseline*

A standard procedure in genetics to analyze candidate genes is to assess the expression level of a candidate gene in patient derived material against wild type. Candidates for which a significant difference is observed between the two groups are considered promising (see for example [20-22]). The higher a candidate's log2 ratio, the higher is its position in the ranking. This simple comparison of log2 ratios is our baseline in this study.

***1. Strategy: Kernel Ridge Regression Ranking*** Consider a given weighted and undirected graph $G$ With symmetric weights $w_{ij} \geq 0$ between nodes $i$ and $j$. The weight $w_{ij}$ increases with the importance of the relation between nodes $i$ and $j$: the larger its value, the easier the communication through the edge. Let $A$ be the Adjacency matrix with $a_{ij} = w_{ij}$ if the nodes $i$ and $j$ are connected and $a_{ij} = 0$ otherwise. The Laplacian matrix $L$ Of $G$ is defined As $L = D - A$ With $D = diag(a_i) = \sum_{j=1}^{n} a_{ij}$. $L$ is symmetric and positive semidefinite.

The *Laplacian Exponential Diffusion Kernel* was introduced by *Kondor and Lafferty (2002)* [44] as

$$K_{LED} = \lim_{n\to\infty}\left(I + \frac{\alpha L}{n}\right)^n = e^{\alpha L} \quad (1)$$

whereby $\alpha$ is the diffusion parameter that determines the degree of diffusion. For a Laplacian matrix, $e^{\alpha L}$ is always positive definite and can thus be used as a kernel matrix. It can be seen as a random walk, starting from a node and transitioning to neighboring nodes with probability $\alpha$. In our application we have applied a negative diffusion parameter, i.e. $K_{LED} = e^{-\alpha L}$.

The *Regularized Laplacian Kernel* was introduced by *Smola & Kondor (2003)* and *Fouss et al. (2006)* [45,46] as

$$K_{RL} = \sum_{k=0}^{\infty} \alpha^k (-L)^k = (I + \alpha L)^{-1}. \quad (2)$$

The *Regularized Commute time Kernel* was introduced by *Fouss et al. 2006* [46] as

$$K_{RCT} = (D - \alpha A)^{-1}. \quad (3)$$

Following the kernel computation, normalization and centering procedures are applied to obtain a valid Similarity Network whose values vary between 0 and 1:

$$\begin{aligned} &\text{Normalization}: K_N = D^* K D^*, D^* = \frac{1}{\sqrt{D}},\ D_{ii} = k(x_i, x_i) \\ &\text{Centering}: K_C = K_N - \frac{1}{n} jj^T K_N - \frac{1}{n} K_N jj^T + \frac{1}{n^2}(j^T K_N j) jj^T \end{aligned} \quad (4)$$

We have defined a neighborhood of a gene as follows. First, it must contain at least one gene other the requested gene. Second, we have limited the amount of neighbors to 50 (see results and parameter tuning in Table S1a (see additional file 1)). Furthermore, we can extract the similarities between genes from the Similarity Network. As a consequence, we obtain for each candidate gene a neighborhood consisting of the 2 to 50 most similar genes in the network that are considered to be influenced in their expression by the candidate gene.

For a given (semi) positive definite kernel matrix $K \in R^n$ and the vector of response variables $Y \in R^n$ , *Saunders et al. (1998)* [47] and *Cawley et al. (2006)* [48] have defined the regression problem $\hat{\gamma} = \sum_{i=1}^{n} a_i K(x_i, x)$ with weights $a \in R^n$ in the Regularization. Theory as

$$\min_a \left\| Y - Ka \right\|_2^2 + \frac{\lambda}{2} a^T a \quad (5)$$

with $\lambda$ as a parameter defining the degree of smoothness in regression. Taking conditions for optimality $Y = (K + \lambda I)a$ , the solution vector $a^*$ is

$$a^* = (K + \lambda I)^{-1} Y. \quad (6)$$

For a new given $x$ it follows

$$\hat{\gamma} = \sum_{i=1}^{n} a_i^* K_x(x_i, x) \quad (7)$$

where $K_x$ is a $n \times 1$ vector containing the kernel evaluations between the candidate point $x$ and $n$ neighboring points (the neighborhood of $x$). The resulting value of $\hat{\gamma}$ can be seen as a *smoothed differential expression value* for the candidate gene $x$ having $n$ neighbors with expression values $Y$.

For each data set all candidates were ranked based on their *smoothed differential expression value*s. A candidate gene with a strongly differentially expressed neighborhood will obtain a *smoothed differential expression value* that is larger than its original differential expression level derived from the microarray experiment. On the other hand, a candidate with a low differentially expressed neighborhood will obtain a small *smoothed differential expression value*, even if its original differential expression level had been large.

***2. Strategy: Heat Kernel Diffusion Ranking*** The *Heat Kernel Diffusion Ranking* approach prioritizes the candidate genes by diffusing the differential expression values of the candidate genes through the network based on the confidence scores of the associations/interactions. We have applied the Heat Kernel rank introduced by [10] and recently used by [25] to unravel relevant regulators in the Saccharomyces cerevisiae regulatory network.

Given a graph $G$, the transition probability matrix $W$ of a random walk on $G$ is defined as $W = D^{-1} A$. Consider $L = I - W$. Given a parameter $\alpha$, establishing the diffusion rate, and a preference vector $p_0$ , expressing the initial relevance score of each node, the ranking $p_\alpha$ is given by

$$p_\alpha = \sum_{k=0}^{\infty} \frac{(-\alpha)^k}{k!} p_0 L^k = p_0 e^{-\alpha L} \quad (8)$$

We have used the discrete approximation by *Yang et al. (2007)*:

$$p_\alpha = p_0 \left(I + \frac{-\alpha}{N} L\right)^N \quad (9)$$

with $N$ being the number of iterations [24]. This iterative diffusion can be regarded as a random walk through

the network which is comparable to the Laplacian Exponential Diffusion Kernel (see equation (1)). However, the Heat Kernel rank considers only an initialized preference vector, whereas the *Kernel Ridge Regression Ranking* uses the Exponential Diffusion Kernel to solve a regression problem.

Following *Francisco et al. (2009)*, performing reduced number of iterations is usually sufficient for ranking purposes using the Heat Kernel rank [25]. In fact, we reach a considerably good performance already after few iterations in our application (see Results section).

We have initialized the preference vector $p_0$ with the differential expression levels coming from a data set from our benchmark. The resulting heat diffusion rank vector $p_\alpha$ contains a score for every candidate gene based on the heat diffusion random walk of its expression level through the network.

***3. Strategy: Arnoldi Diffusion Ranking*** Based on a functional association or PPI network we have computed network diffusion based on a Kyrlov Space method, namely the Arnoldi algorithm, as presented in [11]. The Arnoldi algorithm projects the exponential of a large matrix (here, the *STRING network*, the BioGRID network, or the I2 D network) onto a small Krylov subspace and approximates the matrix exponential operation $e^A v$ as

$$e^A v \approx p_{m-1}(A)v, \tag{10}$$

where $A$ is a matrix of dimension $N$, $v$ is a nonzero vector and $p_{m-1}$is a polynomial of degree *m-1*. Since this approximation is an element of the Krylov subspace $K_m \equiv span\{v, Av, ..., A^{m-1}v\}$, the problem was reformulated by *Saad (1992)* as that of finding an element of $K_m$ that approximates $u = e^A v$ [11]. Based on the Krylov subspace $K_m$ and the Hessenberg matrix $H_m$ of dimension $n \times n$, $u$ can be computed by

$$u = e^A v \approx \|v\| \cdot V_m e^{\alpha H_m} e_1. \tag{11}$$

In this study, we want to compute the *steady state vector of a random walk* based on the *STRING network* which will be an approximation of the Laplacian Exponential Diffusion Kernel presented in equation (1) as $K = e^{\alpha L}$ with $L = I - D^{-1} A$ (see above). The kernel matrix $K$ can be seen as a random walk on the graph with transition probability $\alpha$, and its *ith* column vector represents the steady state probability of a random walk starting at node $i$ with a sufficient number of steps leading to convergence. Then, $K_{ij}$ is the probability of stopping at node $j$ having started at node $i$ after infinite time steps. The column vector $i$ of $K$ can be written as

$$Ke_i = e^{\alpha L} e_i \tag{12}$$

where $e_i$ = [0,...0,1,0,..., 0]' is a zero vector with a 1 at position $i$.

To compute the approximation of the matrix exponential operation $e^A v$ (see equation (12)) considering expression levels for our application, we have initialized the starting vector $v$ with differential expression levels coming from a data set from our benchmark. After running the Arnoldi Algorithm the resulting vector $V_m$ and the corresponding Hessenberg matrix $H_m$ of dimension $m$ (coming from the Krylov subspace) lead to the resulting vector $u = e^A v$ (see equation (13)) reflecting the steady state probability of reaching node $j$ after infinite time steps by starting at the same time point from all nodes that have been initialized in starting vector $v$. The parameter $m$ is the dimension of the Kyrlov subspace and corresponds to the number of iterations in the Arnoldi approximation of network diffusion. The larger $m$, the more accurate is the approximation of network diffusion.

***4. Strategy: Direct Neighborhood Ranking*** In the *Direct Neighborhood Ranking*, we have used a functional association or PPI network directly to capture a neighborhood for a candidate gene without considering any diffusion over the network. A candidate's neighborhood contains all genes $j$ directly connected to the candidate $i$ in the network with weight $w_{ij}$ , representing the probability that an association exists in reality.

Let $x_i$ be the differential expression level of candidate gene $i$, $a$ the weighting parameter that determines the influence of the candidate's expression and its neighborhood's expression, $\varepsilon$ the threshold defining the minimum edge weight between the candidate gene and a neighbor, and $N$ the number of neighboring genes of candidate gene $i$ for which weight $w_{ij}$ is larger then $\varepsilon$. Then the score is computed by

$$\hat{x}_i = a \cdot x_i + (1-a) \cdot \frac{\sum\{j \neq i, j : w_{ij} > \varepsilon\} x_j}{N} \tag{13}$$

## Additional material

**Additional file 1: Supplementary Tables**. This document contains all supplementary tables mentioned in the article.

**Acknowledgements**

We kindly thank Ernesto Iacucci for his helpful review of the manuscript. The research of DN, FO, BDM, and YM was supported by Research Council KUL (ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys, START 1, several PhD/postdoc & fellow grants), Flemish Government (FWO:

PhD/postdoc grants, projects, G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR)), IWT (PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3), FOD (Cancer plans), Belgian Federal Science Policy Office (IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011)), EU-RTD (ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED). The work of JPG was supported by FCT PhD grant SFRH/BD/36586/2007.

**Author details**
[1]Department of Electrical Engineering (ESAT-SCD) Katholieke Universiteit Leuven, 3001 Leuven, Belgium. [2]Knowledge Discovery and Bioinformatics (KDBIO) group, INESC-ID, Rua Alves Redol 9, 1000-029 Lisboa, Portugal. [3]Instituto Superior Técnico, Technical Universityof Lisbon, Av. Rovisco Pais, 1049-001 Lisboa, Portugal.

**Authors' contributions**
DN developed the presented gene prioritization strategies, set up the experiments, analyzed the data and wrote the paper. JPG participated in writing the paper and in developing the presented gene prioritization strategies. FO participated in developing the presented gene prioritization strategies. YM and BDM supervised the project. All authors read and approved the final manuscript.

Received: 27 May 2010 Accepted: 14 September 2010
Published: 14 September 2010

**References**

1. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, *et al*: **Gene prioritization through genomic data fusion.** *Nat Biotech* 2006, **24(5)**:537-44.
2. Köhler S, Bauer S, Horn D, Robinson PN: **Walking the interactome for prioritization of candidate disease genes.** *Am J Hum Genet* 2008, **82(4)**:949-58.
3. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, *et al*: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.** *Am J Hum Genet* 2006, **78(6)**:1011-25.
4. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, *et al*: **A human phenome-interactome network of protein complexes implicated in genetic disorders.** *Nat Biotech* 2007, **25(3)**:309-16.
5. Chen J, Aronow BJ, Jegga AG: **Disease candidate gene identification and prioritization using protein interaction networks.** *BMC Bioinformatics* 2009, **27(10)**:73.
6. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102(43)**:15545-50.
7. Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, *et al*: **Network-Based Analysis of Type 2 Diabetes.** *PLoS Genet* 2007, **15;3(6)**:e96.
8. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, *et al*: **A guide to web tools to prioritize candidate genes.** *Briefings in Bioinformatics* 2010.
9. Nitsch D, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, *et al*: **Network Analysis of Differential Expression for the Identification of Disease-Causing Genes.** *PLoS ONE* 2009, **4(5)**:e5526.
10. Chung F, Yau ST: **Coverings, heat kernels and spanning trees.** *Electronic Journal of Combinatorics* 1999, 6.
11. Saad Y: **Analysis of some Krylov subspace approximations to the matrix exponential operator.** *SIAM Journal on Numerical Analysis (SINUM)* 1992, **29(1)**:209-228.
12. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, *et al*: **NCBI GEO: mining tens of millions of expression profiles - database and tools update.** *Nucleic Acids Res* 2007, , **35 Database:** D760-5.
13. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, *et al*: **STRING 7 - recent developments in the integration and prediction of protein interactions.** *Nucleic Acids Res* 2007, **35**:D358-62.
14. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, *et al*: **STRING 8 - a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, , **37 Database:** D412-6.
15. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, *et al*: **Biogrid: A General Repository for Interaction Datasets.** *Nucleic Acids Res* 2006, **34**: D535-9.
16. Brown KR, Jurisica I: **Online predicted human interaction database.** *Bioinformatics* 2005, **21(9)**:2076-82.
17. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis.** *Mol Syst Biol* 2007, **3**:140.
18. Gupta R, Agrawal S, Rao N, Tian Z, Kuang R, *et al*: **Integrative Biomarker Discovery for Breast Cancer Metastasis from Gene Expression and Protein Interaction Data Using Error-tolerant Pattern Mining.** *Proceedings of the International Conference on Bioinformatics and Computational Biology (BICoB)* 2010.
19. Nibbe RK, Koyutürk M, Chance MR: **An Integrative -omics Approach to Identify Functional Sub-Networks in Human Colorectal Cancer.** *PLoS Comput Biol* 2010, **6(1)**:e1000639.
20. Park JW, Cai J, McIntosh I, Jabs EW, Fallin MD, *et al*: **High throughput SNP and expression analyses of candidate genes for non-syndromic oral clefts.** *J Med Genet* 2006, **43(7)**:598-608.
21. Mirkin S, Arslan M, Churikov D, Corica A, Diaz JI, *et al*: **In search of candidate genes critically expressed in the human endometrium during the window of implantation.** *Hum Reprod* 2005, **20**:2104-2117.
22. Zieker D, Fehrenbach E, Dietzsch J, Fliegner J, Waidmann M, *et al*: **cDNA microarray analysis reveals novel candidate genes expressed in human peripheral blood following exhaustive exercise.** *Physiol Genomics* 2005, **23(3)**:287-94.
23. Baldi P, Long AD: **A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes.** *Bioinformatics* 2001, **17(6)**:509-519.
24. Yang H, King I, Lyu MR: **Diffusionrank: a possible penicillin for web spamming.** In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval: 23-27 July 2007, Amsterdam*. Edited by: Kraaij W, de Vries AP, Clarke CLA, Fuhr N, Kando N. ACM; 2007:431.
25. Francisco AP, Goncalves JP, Madeira SC, Oliveira AL: **Using personalized ranking to unravel relevant regulations in the Saccharomyces cerevisiae regulatory network.** *Proceedings of Jornadas de Bioinformatica* 2009, 3-6 November 2009, Lisbon.
26. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** *Genome Biol* 2005, **6(2)**:R16.
27. Harr B, Schlötterer C: **Comparison of algorithm s for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons.** *Nucleic Acids Res* 2006, **34(2)**:e8.
28. Lim WK, Wang K, Lefebvre C, Califano A: **Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks.** *Bioinformatics* 2007, **23(13)**:i282-8.
29. Vardhanabhuti S, Blakemore SJ, Clark SM, Ghosh S, Stephens RJ, *et al*: **A comparison of statistical tests for detecting differential expression using Affymetrix oligonucleotide microarrays.** *OMICS* 2006, **10(4)**:555-66.
30. Panchenko A, Przytycka T, (eds): *Protein-protein Interactions and Networks* Springer-Verlag, London 2008.
31. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, *et al*: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, , **32 Database:** D449-51.
32. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, *et al*: **BIND-The Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2001, **29(1)**:242-5.
33. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, *et al*: **The IntAct molecular interaction database in 2010.** *Nucleic Acids Res* 2009.
34. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, *et al*: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513(1)**:135-40.
35. Razick S, Magklaras G, Donaldson IM: **iRefIndex: A consolidated protein interaction database with provenance.** *BMC Bioinformatics* 2008, **9(1)**:405.
36. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, *et al*: **The MIPS mammalian protein-protein interaction database.** . *Bioinformatics* 2005, **21(6)**:832-834.
37. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, *et al*: **Human Protein Reference Database - 2009 Update.** *Nucleic Acids Res* **37**: D767-72.

38. **Affymetrix: technical support documentation.** [http://www.affymetrix.com/support/technical/byproduct.affx?product=mas].
39. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, *et al*: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4(2)**:249-64.
40. Wu Z, Irizarry RA: **Stochastic models inspired by hybridization theory for short oligonucleotide arrays.** *J Comput Biol* 2005, **12(6)**:882-93.
41. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy - analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20(3)**:307-15.
42. Fawcett T: **An introduction to ROC analysis.** *Pattern Recognition Letters* 2006, **27**:861-874.
43. Hanley JA, McNeil BJ: **The Meaning and use of the Area under a Receiver Operation Characteristic (ROC) Curve.** *Radiology* 1982, **143**:29-36.
44. Kondor RI, Lafferty JD: **Diffusion kernels on graphs and other discrete structures.** In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML): 8-12 July 2002, Sydney.* Edited by: Sammut C, Hoffmann AG. Morgan Kaufmann; 2002:315-322.
45. Smola AJ, Kondor R: **Kernels and regularization on graphs.** In *Proceedings of the 16th Annual Conference on Learning Theory (COLT): 24-27 August 2003, Washington DC.* Edited by: Schölkopf B, Warmuth M. Springer Verlag; 2003:.
46. Fouss F, Yen L, Pirotte A, Saerens M: **An Experimental Investigation of Graph Kernels on a Collaborative Recommendation Task.** *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM): 18-22 December 2006, Hong Kong* IEEE Computer Society 2006, 863-868.
47. Saunders C, Gammerman A, Vovk V: **Ridge Regression Learning Algorithm in Dual Variables.** In *Proceedings of the 15th International Conference on Machine Learning (ICML): 24-27 July 1998, Madison.* Edited by: Shavlik JW. Morgan Kaufmann Publishers; 1998:.
48. Cawley GC, Talbot NLC, Chapelle O: **Estimating Predictive Variances with Kernel Ridge Regression.** In *Proceedings of Machine Learning Challenges: First PASCAL Machine Learning Challenges Workshop (MLCW): 11-13 July 2005, Southampton.* Edited by: Quinonero-Candela J, Dagan I, Magnini B, D'Alché-Buc F. Springer Verlag; 2006:56-77.

doi:10.1186/1471-2105-11-460
**Cite this article as:** Nitsch *et al.*: Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics* 2010 **11**:460.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

# Supplementary Material

**Supplementary Table S1a:** Parameter optimization for λ and *nn* in the *Kernel Ridge Regression Ranking* using the Laplacian Diffusion Kernel with α=0.5 based on the *STRING network* version 7.1.

| | | | λ= | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0.01 | | | 0.1 | | | 1 | | | 10 | | | 100 | | |
| | | ***nn*** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** |
| | | 30 | 16 | 29 | 0.826 | 21 | 32 | 0.801 | 23 | 32 | 0.839 | 23 | 26 | 0.818 | 21 | 25 | 0.823 |
| | log2 ratio | 50 | 19 | 24 | 0.781 | 20 | 32 | 0.843 | 23 | 29 | 0.809 | 21 | 25 | 0.759 | 22 | 25 | 0.768 |
| | | 100 | 15 | 23 | 0.795 | 17 | 26 | 0.776 | 19 | 26 | 0.793 | 18 | 24 | 0.748 | 18 | 21 | 0.748 |
| | | 30 | 22 | 31 | 0.848 | 21 | 31 | 0.826 | 25 | 31 | 0.853 | 23 | 27 | 0.740 | 25 | 29 | 0.758 |
| **RMA** | sign. log2 ratio | 50 | 25 | 27 | 0.821 | 23 | 31 | 0.868 | 27 | 32 | 0.868 | 22 | 24 | 0.731 | 21 | 25 | 0.701 |
| | | 100 | 18 | 27 | 0.798 | 17 | 27 | 0.804 | 22 | 28 | 0.828 | 21 | 26 | 0.727 | 20 | 24 | 0.691 |
| | | 30 | 17 | 26 | 0.780 | 20 | 26 | 0.761 | 23 | 26 | 0.789 | 18 | 22 | 0.694 | 16 | 21 | 0.674 |
| | test statistic | 50 | 18 | 23 | 0.780 | 19 | 26 | 0.797 | 20 | 26 | 0.771 | 16 | 21 | 0.683 | 13 | 19 | 0.683 |
| | | 100 | 15 | 25 | 0.757 | 19 | 25 | 0.788 | 18 | 28 | 0.786 | 18 | 24 | 0.735 | 14 | 20 | 0.719 |
| | | 30 | 20 | 26 | 0.820 | 24 | 32 | 0.837 | 25 | 30 | 0.830 | 24 | 26 | 0.766 | 22 | 23 | 0.754 |
| | log2 ratio | 50 | 20 | 27 | 0.843 | 21 | 28 | 0.820 | 21 | 28 | 0.769 | 21 | 24 | 0.698 | 19 | 21 | 0.686 |
| | | 100 | 16 | 21 | 0.787 | 15 | 27 | 0.829 | 20 | 28 | 0.818 | 20 | 21 | 0.681 | 17 | 21 | 0.683 |
| | | 30 | 22 | 32 | 0.864 | 22 | 31 | 0.860 | 24 | 33 | 0.886 | 24 | 25 | 0.729 | 23 | 26 | 0.701 |
| **GCRMA** | sign. log2 ratio | 50 | 20 | 29 | 0.814 | 22 | 29 | 0.819 | 27 | 31 | 0.835 | 23 | 26 | 0.715 | 20 | 24 | 0.706 |
| | | 100 | 14 | 25 | 0.780 | 17 | 26 | 0.805 | 21 | 29 | 0.778 | 21 | 24 | 0.744 | 21 | 24 | 0.693 |
| | | 30 | 17 | 26 | 0.791 | 21 | 28 | 0.791 | 20 | 27 | 0.788 | 16 | 24 | 0.717 | 15 | 23 | 0.711 |
| | test statistic | 50 | 19 | 24 | 0.792 | 19 | 25 | 0.760 | 19 | 22 | 0.745 | 16 | 24 | 0.727 | 16 | 21 | 0.694 |
| | | 100 | 16 | 23 | 0.791 | 17 | 26 | 0.778 | 18 | 23 | 0.735 | 15 | 21 | 0.653 | 14 | 22 | 0.711 |
| | | 30 | 16 | 26 | 0.758 | 16 | 24 | 0.749 | 12 | 17 | 0.695 | 17 | 20 | 0.763 | 17 | 22 | 0.770 |
| | log2 ratio | 50 | 15 | 25 | 0.767 | 21 | 29 | 0.824 | 18 | 24 | 0.736 | 16 | 21 | 0.749 | 16 | 23 | 0.737 |
| | | 100 | 15 | 21 | 0.768 | 16 | 23 | 0.766 | 19 | 26 | 0.809 | 11 | 14 | 0.723 | 13 | 21 | 0.688 |
| | | 30 | 21 | 28 | 0.823 | 23 | 30 | 0.802 | 21 | 28 | 0.798 | 20 | 26 | 0.706 | 20 | 25 | 0.724 |
| **MAS5** | sign. log2 ratio | 50 | 17 | 29 | 0.821 | 22 | 32 | 0.833 | 23 | 29 | 0.834 | 19 | 28 | 0.740 | 23 | 24 | 0.747 |
| | | 100 | 16 | 26 | 0.794 | 20 | 28 | 0.785 | 25 | 29 | 0.819 | 16 | 23 | 0.689 | 21 | 24 | 0.714 |
| | | 30 | 17 | 27 | 0.781 | 15 | 28 | 0.772 | 22 | 27 | 0.757 | 18 | 23 | 0.742 | 18 | 22 | 0.739 |
| | test statistic | 50 | 19 | 24 | 0.759 | 19 | 22 | 0.734 | 13 | 29 | 0.790 | 14 | 23 | 0.726 | 13 | 22 | 0.734 |
| | | 100 | 12 | 21 | 0.762 | 15 | 25 | 0.766 | 13 | 18 | 0.717 | 16 | 21 | 0.750 | 15 | 19 | 0.693 |

**Supplementary Table S1b:** Prioritization results based on different kernels based on equation (1)-(3) and on the *STRING network* version 7.1.

| | | $K_{LED}, \alpha = 0.5$ | | | $K_{RCT}, \alpha = 0.90$ | | | $K_{RCT}, \alpha = 0.95$ | | | $K_{RL}, \alpha = 1$ | | | $K_{RL}, \alpha = 2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** |
| | log2 ratio | 23 | 29 | 0.809 | 16 | 22 | 0.767 | 18 | 25 | 0.777 | 21 | 24 | 0.784 | 20 | 28 | 0.789 |
| **RMA** | sign. log2 ratio | 27 | 32 | 0.868 | 22 | 26 | 0.766 | 24 | 26 | 0.774 | 24 | 27 | 0.757 | 21 | 26 | 0.799 |
| | test statistic | 20 | 26 | 0.771 | 16 | 20 | 0.693 | 16 | 20 | 0.710 | 14 | 21 | 0.685 | 14 | 20 | 0.698 |
| | log2 ratio | 21 | 28 | 0.769 | 20 | 22 | 0.755 | 20 | 23 | 0.763 | 21 | 23 | 0.767 | 23 | 26 | 0.792 |
| **GCRMA** | sign. log2 ratio | 27 | 31 | 0.835 | 22 | 25 | 0.782 | 22 | 29 | 0.792 | 25 | 28 | 0.765 | 24 | 25 | 0.772 |
| | test statistic | 19 | 22 | 0.745 | 19 | 2 | 0.670 | 18 | 21 | 0.706 | 20 | 21 | 0.664 | 16 | 20 | 0.677 |
| | log2 ratio | 18 | 24 | 0.736 | 11 | 17 | 0.666 | 10 | 16 | 0.668 | 12 | 15 | 0.657 | 12 | 19 | 0.691 |
| **MAS5** | sign. log2 ratio | 23 | 29 | 0.834 | 21 | 25 | 0.749 | 21 | 25 | 0.761 | 19 | 25 | 0.721 | 19 | 27 | 0.754 |
| | test statistic | 13 | 29 | 0.790 | 14 | 18 | 0.656 | 14 | 21 | 0.689 | 13 | 21 | 0.674 | 11 | 20 | 0.685 |

**Supplementary Table S1c:** Parameter optimization for *m* in the *Heat Kernel Diffusion Ranking* based on the *STRING network* version 7.1.

| | | ***m*=** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | | | **2** | | | **3** | | | **5** | | | **10** | | |
| | | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **Top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** |
| **RMA** | log2 ratio | 32 | 34 | 0.900 | 32 | 35 | 0.902 | 32 | 34 | 0.902 | 32 | 34 | 0.902 | 32 | 34 | 0.902 |
| | sign. log2 ratio | 31 | 34 | 0.910 | 31 | 34 | 0.908 | 31 | 34 | 0.907 | 31 | 34 | 0.909 | 31 | 34 | 0.909 |
| | test statistic | 32 | 34 | 0.901 | 32 | 34 | 0.900 | 32 | 34 | 0.900 | 32 | 33 | 0.899 | 32 | 33 | 0.900 |
| **GCRMA** | log2 ratio | 31 | 33 | 0.905 | 31 | 33 | 0.906 | 31 | 33 | 0.906 | 31 | 33 | 0.905 | 31 | 33 | 0.905 |
| | sign. log2 ratio | 28 | 33 | 0.889 | 28 | 34 | 0.898 | 28 | 34 | 0.899 | 28 | 34 | 0.899 | 28 | 34 | 0.899 |
| | test statistic | 32 | 33 | 0.895 | 32 | 33 | 0.897 | 32 | 33 | 0.897 | 32 | 33 | 0.897 | 32 | 33 | 0.897 |
| **MAS5** | log2 ratio | 26 | 32 | 0.877 | 28 | 31 | 0.876 | 28 | 31 | 0.875 | 28 | 31 | 0.877 | 28 | 31 | 0.878 |
| | sign. log2 ratio | 26 | 31 | 0.877 | 26 | 31 | 0.888 | 26 | 31 | 0.887 | 26 | 31 | 0.888 | 26 | 31 | 0.888 |
| | test statistic | 32 | 32 | 0.890 | 31 | 32 | 0.888 | 31 | 32 | 0.889 | 31 | 32 | 0.889 | 31 | 32 | 0.889 |

**Supplementary Table S1d:** Prioritization results based on different initializations of the preference vector in the *Heat Kernel Diffusion Ranking* based on the *STRING network* version 8.2. The first three columns show the results after initializing the preference vector with only the expression values of the candidate genes, all other genes were set to *0*. The last three columns show the results after initializing the preference vector with all expression values available from the data set.

| | | **use only candidate's expression** | | | **differentially expressed=1, otherwise =0** | | | **all originl expression** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **top 10** | **top 20** | **AUC** | **top 10** | **top 20** | **AUC** | **top 10** | **top 20** | **AUC** |
| **RMA** | log2 ratio | 27 | 30 | 0.845 | 20 | 26 | 0.814 | 32 | 35 | 0.902 |
| | sign. log2 ratio | 30 | 32 | 0.903 | 29 | 33 | 0.902 | 31 | 34 | 0.908 |
| | test statistic | 27 | 30 | 0.843 | 20 | 26 | 0.814 | 32 | 34 | 0.9 |
| **GCRMA** | log2 ratio | 25 | 31 | 0.865 | 23 | 26 | 0.831 | 31 | 33 | 0.906 |
| | sign. log2 ratio | 26 | 31 | 0.89 | 26 | 33 | 0.893 | 28 | 34 | 0.898 |
| | test statistic | 27 | 31 | 0.855 | 23 | 26 | 0.831 | 32 | 33 | 0.897 |
| **MAS5** | log2 ratio | 25 | 27 | 0.84 | 21 | 26 | 0.814 | 28 | 31 | 0.876 |
| | sign. log2 ratio | 27 | 30 | 0.888 | 24 | 32 | 0.889 | 26 | 31 | 0.888 |
| | test statistic | 25 | 28 | 0.846 | 21 | 26 | 0.814 | 31 | 32 | 0.888 |
| **binary values, no expression** | | 5 | 12 | 0.613 | | | | | | |

**Supplementary Table S1e:** Parameter optimization for ε in the *Direct Neighborhood Ranking* based on the *STRING network* version 7.1.

| | | **ε =** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **0.15** | | | **0.5** | | | **0.7** | | |
| | | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** |
| **RMA** | log2 ratio | 27 | 31 | 0.859 | 1 | 10 | 0.671 | 2 | 3 | 0.476 |
| | sign. log2 ratio | 28 | 31 | 0.880 | 1 | 10 | 0.702 | 2 | 3 | 0.471 |
| | test statistic | 29 | 30 | 0.856 | 1 | 10 | 0.669 | 2 | 3 | 0.472 |
| **GCRMA** | log2 ratio | 27 | 31 | 0.874 | 1 | 7 | 0.689 | 2 | 3 | 0.481 |
| | sign. log2 ratio | 25 | 28 | 0.855 | 1 | 8 | 0.689 | 2 | 3 | 0.474 |
| | test statistic | 27 | 31 | 0.863 | 1 | 7 | 0.681 | 2 | 3 | 0.472 |
| **MAS5** | log2 ratio | 23 | 27 | 0.846 | 1 | 8 | 0.666 | 2 | 3 | 0.463 |
| | sign. log2 ratio | 25 | 28 | 0.844 | 1 | 10 | 0.682 | 2 | 3 | 0.473 |
| | test statistic | 27 | 30 | 0.855 | 1 | 10 | 0.677 | 2 | 3 | 0.476 |

**Supplementary Table S1f:** Parameter optimization for *m* in the *Arnoldi Diffusion Ranking* based on the *STRING network* version 7.1.

| | | ***m*=** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | | | **2** | | | **10** | | | **50** | | |
| | | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** | **top10** | **top20** | **AUC** |
| **RMA** | log2 ratio | 2 | 7 | 0.419 | 27 | 31 | 0.857 | 27 | 31 | 0.857 | 27 | 31 | 0.857 |
| | sign. log2 ratio | 2 | 7 | 0.419 | 28 | 31 | 0.885 | 28 | 31 | 0.887 | 28 | 31 | 0.887 |
| | test statistic | 2 | 7 | 0.419 | 28 | 30 | 0.855 | 28 | 30 | 0.855 | 28 | 30 | 0.855 |
| **GCRMA** | log2 ratio | 2 | 7 | 0.419 | 27 | 32 | 0.875 | 27 | 31 | 0.876 | 27 | 31 | 0.876 |
| | sign. log2 ratio | 2 | 7 | 0.419 | 25 | 29 | 0.860 | 25 | 30 | 0.871 | 25 | 30 | 0.871 |
| | test statistic | 2 | 7 | 0.419 | 27 | 31 | 0.865 | 27 | 31 | 0.865 | 27 | 31 | 0.865 |
| **MAS5** | log2 ratio | 2 | 7 | 0.419 | 25 | 27 | 0.849 | 25 | 27 | 0.849 | 25 | 27 | 0.849 |
| | sign. log2 ratio | 2 | 7 | 0.419 | 25 | 29 | 0.853 | 25 | 29 | 0.864 | 25 | 29 | 0.864 |
| | test statistic | 2 | 7 | 0.419 | 26 | 30 | 0.858 | 26 | 30 | 0.860 | 26 | 30 | 0.860 |

**Supplementary Table S2:** Analysis of dependency on expression levels: comparison of performance of the four strategies using the original expression level of the knockout gene and 0 as its expression level based on the *STRING network* version 7.1

| | | | **Original expression levels** | | | **expression (KO) = 0** | | |
|---|---|---|---|---|---|---|---|---|
| | | | **top 10** | **top 20** | **AUC** | **top 10** | **top 20** | **AUC** |
| **RMA** | **Standard genetic procedure: Simple Expression Ranking** | | **20** | **25** | **0.801** | **0** | **0** | **na** |
| | **Direct Neighborhood Ranking** $\varepsilon$>0.15, *a*=0.5 | log2 ratio | 27 | 31 | 0.859 | 0 | 0 | 0.363 |
| | | sign. log2 ratio | 28 | 31 | 0.880 | 5 | 8 | 0.673 |
| | | test statistic | 29 | 30 | 0.856 | 0 | 0 | 0.352 |
| | **Kernel Ridge Regression Ranking** $\lambda$=1, *nn=50*, $K_{LED}, \alpha = 0.5$ | log2 ratio | 23 | 29 | 0.809 | 9 | 10 | 0.597 |
| | | sign. log2 ratio | 27 | 32 | 0.868 | 8 | 11 | 0.494 |
| | | test statistic | 20 | 26 | 0.771 | 7 | 9 | 0.508 |
| | **Heat Kernel Diffusion Ranking** all expression values for $p_0$, *m*=2, α=0.5 | log2 ratio | 32 | 34 | 0.900 | 6 | 8 | 0.529 |
| | | sign. log2 ratio | 31 | 34 | 0.910 | 8 | 14 | 0.735 |
| | | test statistic | 32 | 34 | 0.901 | 6 | 8 | 0.522 |
| | **Arnoldi Diffusion Ranking** *m*=2, α=0.5 | log2 ratio | 27 | 31 | 0.857 | 0 | 0 | 0.312 |
| | | sign. log2 ratio | 28 | 31 | 0.885 | 4 | 10 | 0.685 |
| | | test statistic | 28 | 30 | 0.855 | 0 | 0 | 0.308 |
| **GCRMA** | **Standard genetic procedure: Simple Expression Ranking** | | **18** | **24** | **0.777** | **0** | **0** | **na** |
| | **Direct Neighborhood Ranking** $\varepsilon$>0.15, *a*=0.5 | log2 ratio | 27 | 31 | 0.874 | 0 | 3 | 0.487 |
| | | sign. log2 ratio | 25 | 28 | 0.855 | 4 | 8 | 0.666 |
| | | test statistic | 27 | 31 | 0.863 | 0 | 1 | 0.437 |
| | **Kernel Ridge Regression Ranking** $\lambda$=1, *nn=50*, $K_{LED}, \alpha = 0.5$ | log2 ratio | 21 | 28 | 0.769 | 9 | 10 | 0.544 |
| | | sign. log2 ratio | 27 | 31 | 0.835 | 7 | 8 | 0.479 |
| | | test statistic | 19 | 22 | 0.745 | 7 | 1 | 0.492 |
| | **Heat Kernel Diffusion Ranking** all expression values for $p_0$, *m*=2, α=0.5 | log2 ratio | 31 | 33 | 0.905 | 5 | 10 | 0.598 |
| | | sign. log2 ratio | 28 | 33 | 0.889 | 9 | 19 | 0.741 |
| | | test statistic | 32 | 33 | 0.895 | 6 | 8 | 0.564 |
| | **Arnoldi Diffusion Ranking** *m*=2, α=0.5 | log2 ratio | 27 | 32 | 0.875 | 0 | 1 | 0.453 |
| | | sign. log2 ratio | 25 | 29 | 0.860 | 3 | 10 | 0.681 |
| | | test statistic | 27 | 31 | 0.865 | 0 | 1 | 0.406 |
| **MAS5** | **Standard genetic procedure: Simple Expression Ranking** | | **24** | **28** | **0.837** | **0** | **0** | **na** |
| | **Direct Neighborhood Ranking** $\varepsilon$>0.15, *a*=0.5 | log2 ratio | 23 | 27 | 0.846 | 0 | 0 | 0.340 |
| | | sign. log2 ratio | 25 | 28 | 0.844 | 3 | 8 | 0.651 |
| | | test statistic | 27 | 30 | 0.855 | 0 | 0 | 0.323 |
| | **Kernel Ridge Regression Ranking** $\lambda$=1, *nn=50*, $K_{LED}, \alpha = 0.5$ | log2 ratio | 18 | 24 | 0.736 | 9 | 12 | 0.606 |
| | | sign. log2 ratio | 23 | 29 | 0.834 | 8 | 9 | 0.485 |
| | | test statistic | 13 | 18 | 0.790 | 6 | 10 | 0.476 |
| | **Heat Kernel Diffusion Ranking** all expression values for $p_0$, *m*=2, α=0.5 | log2 ratio | 26 | 32 | 0.877 | 4 | 8 | 0.521 |
| | | sign. log2 ratio | 26 | 31 | 0.877 | 8 | 16 | 0.735 |
| | | test statistic | 32 | 32 | 0.890 | 6 | 8 | 0.516 |
| | **Arnoldi Diffusion Ranking** *m*=2, α=0.5 | log2 ratio | 25 | 27 | 0.849 | 0 | 0 | 0.291 |
| | | sign. log2 ratio | 25 | 29 | 0.853 | 3 | 12 | 0.674 |
| | | test statistic | 26 | 30 | 0.858 | 0 | 0 | 0.289 |

**Supplementary Table S3:** Overview of prioritization results based on optimized parameter settings for all presented strategies using the I2D network.

| | | | top 10% | top 20% | AUC |
|---|---|---|---|---|---|
| **RMA** | **Standard genetic procedure: Simple Expression Ranking** | | **20** | **25** | **0.801** |
| | **Direct Neighborhood Ranking** $\varepsilon$>0.15, $a$=0.5 | log2 ratio | 1 | 6 | 0.394 |
| | | sign. log2 ratio | 1 | 6 | 0.394 |
| | | test statistic | 1 | 6 | 0.394 |
| | **Kernel Ridge Regression Ranking** $\lambda$=1, $nn$=50, $K_{LED}, \alpha = 0.5$ | log2 ratio | 9 | 13 | 0.600 |
| | | sign. log2 ratio | 13 | 13 | 0.587 |
| | | test statistic | 8 | 13 | 0.608 |
| | **Heat Kernel Diffusion Ranking** all expression values for $p_0$, $m$=2, α=0.5 | log2 ratio | 23 | 27 | **0.805** |
| | | sign. log2 ratio | 23 | 29 | 0.790 |
| | | test statistic | 4 | 6 | 0.313 |
| | **Arnoldi Diffusion Ranking** $m$=2, α=0.5 | log2 ratio | 23 | 25 | 0.776 |
| | | sign. log2 ratio | 23 | 26 | 0.755 |
| | | test statistic | 4 | 5 | 0.302 |
| **GCRMA** | **Standard genetic procedure: Simple Expression Ranking** | | **18** | **24** | **0.777** |
| | **Direct Neighborhood Ranking** $\varepsilon$>0.15, $a$=0.5 | log2 ratio | 1 | 6 | 0.394 |
| | | sign. log2 ratio | 1 | 6 | 0.394 |
| | | test statistic | 1 | 6 | 0.394 |
| | **Kernel Ridge Regression Ranking** $\lambda$=1, $nn$=50, $K_{LED}, \alpha = 0.5$ | log2 ratio | 7 | 12 | 0.543 |
| | | sign. log2 ratio | 7 | 10 | 0.615 |
| | | test statistic | 9 | 14 | 0.606 |
| | **Heat Kernel Diffusion Ranking** all expression values for $p_0$, $m$=2, α=0.5 | log2 ratio | 20 | 24 | **0.786** |
| | | sign. log2 ratio | 20 | 24 | **0.797** |
| | | test statistic | 4 | 5 | 0.306 |
| | **Arnoldi Diffusion Ranking** $m$=2, α=0.5 | log2 ratio | 20 | 26 | **0.786** |
| | | sign. log2 ratio | 20 | 23 | 0.745 |
| | | test statistic | 4 | 5 | 0.289 |
| **MAS5** | **Standard genetic procedure: Simple Expression Ranking** | | **24** | **28** | **0.837** |
| | **Direct Neighborhood Ranking** $\varepsilon$>0.15, $a$=0.5 | log2 ratio | 1 | 6 | 0.394 |
| | | sign. log2 ratio | 1 | 6 | 0.394 |
| | | test statistic | 1 | 6 | 0.394 |
| | **Kernel Ridge Regression Ranking** $\lambda$=1, $nn$=50, $K_{LED}, \alpha = 0.5$ | log2 ratio | 3 | 8 | 0.589 |
| | | sign. log2 ratio | 7 | 18 | 0.675 |
| | | test statistic | 7 | 10 | 0.604 |
| | **Heat Kernel Diffusion Ranking** all expression values for $p_0$, $m$=2, α=0.5 | log2 ratio | 17 | 25 | 0.769 |
| | | sign. log2 ratio | 21 | 27 | 0.796 |
| | | test statistic | 3 | 3 | 0.276 |
| | **Arnoldi Diffusion Ranking** $m$=2, α=0.5 | log2 ratio | 15 | 25 | 0.775 |
| | | sign. log2 ratio | 21 | 25 | 0.782 |
| | | test statistic | 3 | 3 | 0.269 |

# Chapter 6

# *Study 5*: PINTA - a web server for network-based gene prioritization from expression data

In this study we have developed a freely accessible web server (PINTA) that is based on the algorithmic implementation of *Study 3* and *Study 4* that makes our method publically available for the biological and medical researchers, aiming at identifying novel disease genes using disease specific expression data. PINTA supports both candidate gene prioritization as well as genome-wide gene prioritization and is available for distinct species. As input data, PINTA only requires disease specific expression data, from which PINTA computes a gene ranking list which is presented to the user and which can be downloaded for further studies.

**Publication**
This work has been published as an electronic version in the *Nucleic Acids Research* journal (ISI IF 2009: 7.4).
*Reference:* Nitsch D, Tranchevent L, Goncalves JP, Vogt JK, Madeira SC, Moreau Y (2011). PINTA - A web server for network-based gene prioritization from expression data. *Nucleic Acids Research* vol. 39 (suppl 2): W334-W338.

**Personal contribution**
The PhD candidate designed and implemented the web server and drafted the paper. The first co-author participated in writing the paper and the third co-author supported the PhD candidate in implementing the web server.

Nucleic Acids Research Advance Access published May 20, 2011

*Nucleic Acids Research, 2011, 1–5*
*doi:10.1093/nar/gkr289*

# PINTA: a web server for network-based gene prioritization from expression data

**Daniela Nitsch[1], Léon-Charles Tranchevent[1], Joana P. Gonçalves[2,3], Josef Korbinian Vogt[4], Sara C. Madeira[2,3] and Yves Moreau[1,*]**

[1]Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, 3001 Leuven, Belgium, [2]Knowledge Discovery and Bioinformatics group (KDBIO), INESC-ID, 1000-029 Lisbon, Portugal, [3]Instituto Superior Técnico (IST), Technical University of Lisbon, 1049-001 Lisbon, Portugal and [4]Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, 2800 Lyngby, Denmark

Received January 31, 2011; Revised March 15, 2011; Accepted April 12, 2011

## ABSTRACT

**PINTA (available at http://www.esat.kuleuven.be/pinta/; this web site is free and open to all users and there is no login requirement) is a web resource for the prioritization of candidate genes based on the differential expression of their neighborhood in a genome-wide protein–protein interaction network. Our strategy is meant for biological and medical researchers aiming at identifying novel disease genes using disease specific expression data. PINTA supports both candidate gene prioritization (starting from a user defined set of candidate genes) as well as genome-wide gene prioritization and is available for five species (human, mouse, rat, worm and yeast). As input data, PINTA only requires disease specific expression data, whereas various platforms (e.g. Affymetrix) are supported. As a result, PINTA computes a gene ranking and presents the results as a table that can easily be browsed and downloaded by the user.**

## BACKGROUND

A major challenge in human genetics is to identify novel disease genes to understand the mechanisms underlying genetic conditions and, in the long term, elaborate novel treatments for these disorders. Genetic studies, such as association studies and linkage analyses, identify chromosomal regions involved in a disease or phenotype of interest, but often result in large lists of candidate genes of which only one or a few are really associated to the disease or phenotype under study. Identifying, among such a list, the most promising candidate genes for a disease of interest has been defined as the gene prioritization problem. Candidate gene prioritization is key in genetics because it is generally too expensive and time-consuming to experimentally validate all candidate genes. Because of the huge amount of genomic data that is publicly available, computational approaches have been developed to avoid performing candidate gene prioritization manually.

In the past couple of years, several gene prioritization methods have been proposed by the bioinformatics community to address this problem. For a detailed review of web based gene prioritization tools and their information sources, the reader is referred to our recent review (1) and its associated web site (http://www.esat.kuleuven.be/gpp). Most of the available gene prioritization tools combine different data and information sources, among which the most commonly used data sources are literature, functional annotations, interactions, expression data and sequence information (2,3). Among these tools, ToppGene (4), SNPs3D (5), GeneDistiller (6) and Posmed (7) additionally include model organism data (mainly mouse data). Other tools, such as GeneWanderer (8), Prioritizer (9) and PhenoPred (10) make use exclusively of genome-wide protein–protein interaction networks. Prioritizer integrates several networks obtained from different databases (including expression data) and uses this huge network to investigate diseases for which several loci are known. Candidate genes from one locus that are connected to candidate genes in another locus are considered promising candidate genes. GeneWanderer uses a global network distance measure to define similarity in protein–protein interaction networks. PhenoPred is based on a human protein–protein interaction network and uses a supervised algorithm for detecting gene-disease associations, known gene-disease associations, protein sequence and protein functional information.

However, most of these tools are using a guilt-by-association concept (candidate genes that are similar

*To whom correspondence should be addressed. Tel: +32 (0)16 32 10 75; Fax: +32 (0)16 32 19 70; Email: yves.moreau@esat.kuleuven.be

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from nar.oxfordjournals.org at KU Leuven on May 23, 2011

to the already confirmed disease genes are considered promising) and are therefore not applicable when little is known about the phenotype or when no confirmed disease genes are available beforehand. One common strategy to circumvent that problem is to rely on keywords to define the genetic condition under study. However, most of the existing tools that accept keywords also rely solely on text-mining of the literature and are therefore less suitable for novel discoveries (11–16). In a recent study (17,18), we have proposed a method that overcomes this limitation by representing prior knowledge about the biological process by experimental data on differential gene expression between affected and healthy individuals.

At the core of the method are a protein–protein interaction or association network and a disease specific expression data set. The method propagates the expression data over the network using an extended random walk approach. Candidate genes are then ranked based on the differential expression level (e.g. fold changes from a case–control study) of their neighborhood. Our method relies on the assumption that strong candidate genes tend to be surrounded by many differentially expressed genes in a genome-wide protein–protein interaction network. This allows the detection of a strong signal for a candidate even if its own differential expression value is too small to be detected by a standard analysis, as long as its interacting partners are highly differentially expressed. Our benchmark on 40 publicly available knockout experiments in mice showed that it outperforms a standard procedure in genetics that ranks candidate genes based solely on their own expression levels (18).

In this article, we describe a novel web server called PINTA (http://www.esat.kuleuven.be/pinta/) that implements the method we have developed previously (17,18). PINTA is a free, user-friendly and easy accessible web tool, which performs candidate gene prioritization (i.e. starting from a predefined set of candidate genes) as well as genome-wide gene prioritization using the method described above. To our knowledge, PINTA is the first web based tool that can prioritize candidate genes for diseases with only limited information about the phenotype or no confirmed disease genes by replacing this knowledge by expression data to model the disease under study. To be of use to a large range of biologists and geneticists, we have made PINTA available for multiple species besides human (mouse, rat, worm and yeast) that represent some of the most common model organisms for human. In addition to the approaches described in (17,18), PINTA can also propagate the expression over the network based on other extended random walk approaches [HITS with priors and k-step Markov (19)]. PINTA also supports the use of probe set names from various Affymetrix platforms.

## PINTA WORKFLOW

A four-step wizard guides the user through the gene prioritization procedure of PINTA (Figure 1). In the first step, the user selects the organism of interest among human, mouse, rat, worm or yeast. In the second step, the user defines whether PINTA needs to compute a genome-wide ranking or to rank a set of candidate genes. In the latter case, the user can choose to provide candidate genes using either official gene symbols or Ensembl identifiers. In the third step, the user can optionally choose whether he wants PINTA to perform a comparison analysis between its own ranking and the ranking obtained by simply ordering the differential expression values. The fourth step consists in uploading disease specific expression data, where probe set names of various Affymetrix platforms are also supported (see below). PINTA uses by default the best performing settings that were determined in our benchmark (18), which is convenient for non-expert users. Advanced users can fine tune the parameters to fit their needs and make the best of their data. Users can choose between five different prioritization algorithms [Heat Kernel Ranking (18), Arnoldi Diffusion Ranking (18), Random Walk on a graph (20), HITS with priors (19) or k-step Markov (19), for more details see below] and two different networks [STRING version 8.2 (21) and I2D version 1.8 (22)]. By default, the Heat Kernel Ranking and the STRING network are used.

## RANKING WITH DIFFERENTIAL EXPRESSION DATA

As input data, PINTA only requires disease-specific expression data in the form of a text file in which each gene is assessed by a differential expression value. The user has to upload a tab delimited text file containing the expression data in the form of two columns: the first column contains the gene identifiers (Affymetrix probe identifiers, gene symbols or Ensembl identifiers) and the second column contains the gene's differential expression signal.

PINTA supports some of the most common Affymetrix chips to accommodate the user with an easier handling of the data and to avoid the need for manual mapping of probe id to gene: HG 1.0 ST, HEx 1.0 ST, HGU 133 Plus2, MouseGenome 430 2.0, MouseGene 1.0 ST, MouseExon 1.0 ST, RatExpressionArray 230 2.0, RatGene 1.0 ST, RatExon 1.0 ST, *Caenorhabditis elegans* Genome Array and Yeast 2.0 Array. If the user is not using one of these Affymetrix chips, he can still upload the expression data using Ensembl identifiers or gene symbols according to the drop down menu of PINTA.

PINTA requires one differential expression value per gene and is neither performing any kind of preprocessing nor normalization of the data; this step lies in the responsibility of the user. If the microarray experiment consists of several experiments or chips, the user has to provide one characteristic expression value for each gene/probe (e.g. the average expression value across all chips). The text file can be created in Microsoft® Excel and then converted into a tab delimited text file, or it can be directly created as a text file as an output of any preprocessing or normalization tool. Since the text file may contain headers

Downloaded from nar.oxfordjournals.org at KU Leuven on May 23, 2011

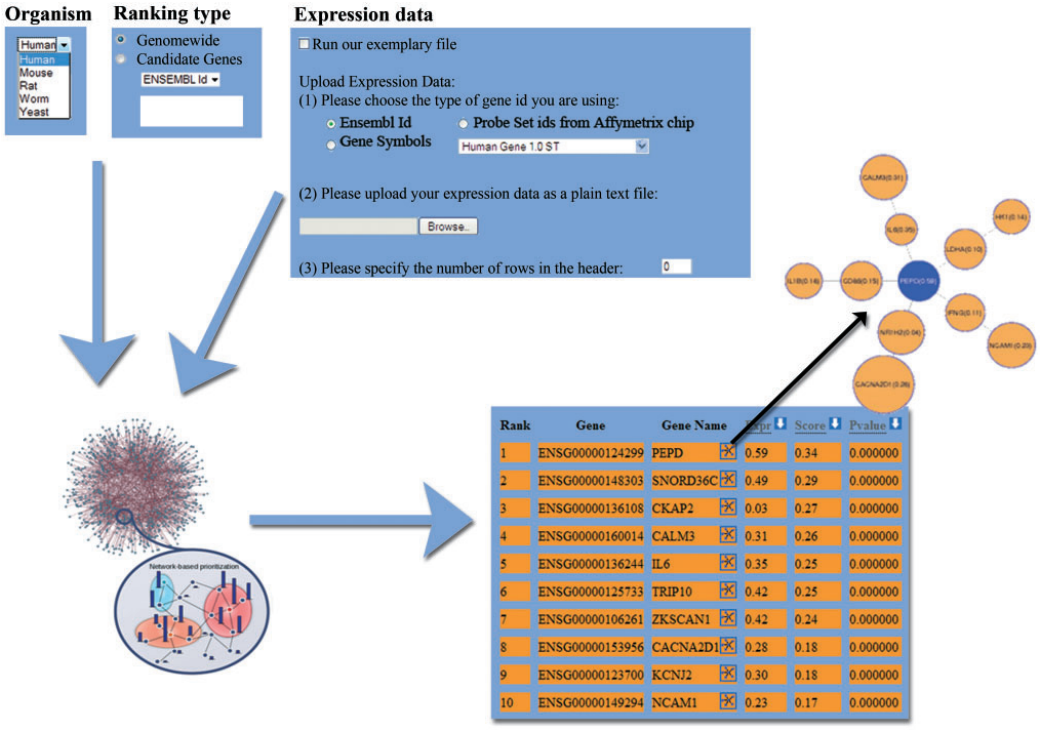



**Figure 1.** The PINTA workflow. In the first step, the user selects the organism of interest. The second step consists in defining whether PINTA needs to compute a genome-wide ranking or to rank a defined set of candidate genes. In the last step, a disease-specific expression data is required in the form of a text file in which each gene is assessed by an expression value. PINTA computes a ranking of all candidate genes and presents the results in a table containing the gene ranks, internal scores, *P*-values and Bonferroni–Holm *P*-values (multiple testing correction). For the top ranked genes, PINTA provides (with the default settings) a graphical view of the strongest contributing interacting genes that lead to the candidate's strong scoring signal together with their expression signal in the network.

(e.g. internal describable notes), the user can specify how many rows of the file belong to the header and PINTA will ignore these rows in its internal computation. If the user is unsure how to choose the correct format for the text file, he can go back to the exemplary text file provided online by PINTA.

PINTA provides five different prioritization algorithms from which the user can choose in the advanced setting by activating the corresponding radio button. Performing a random walk on a network consists in taking successive random steps (20). In the Heat Kernel Ranking (18) (default algorithm in PINTA), the input signal given by the expression level of a candidate is adjusted according to a global network measure taking direct and indirect association into account, whereas the Arnoldi Diffusion Ranking performs network diffusion by applying the Arnoldi algorithm (18,23). The HITS with priors algorithm (19) is a random walk on the network based on mutual reinforcement relation between authorities and hubs using prior knowledge about the importance of nodes and the k-step Markov algorithm (19) is a random walk on the network defined by the probability transition of the graph.

## PRIORITIZATION RESULTS AND DATA VISUALIZATION

PINTA computes a ranking of the candidate genes and presents the results in a table containing the gene ranks, internal scores, *P*-values and Bonferroni–Holm *P*-values (multiple testing correction). The *P*-values are calculated by random permutation on the expression data (random reassignment of the expression values to network nodes and computation of the corresponding randomized scores for all candidate genes). The user can sort this table by *P*-values, scores or expression values to check the differences in the corresponding rankings. PINTA provides for the top ranked genes (with the default settings) a graphical view of the strongest contributing interacting genes and their expression signal in the network. With this graphical view, the user can assess the importance of a top-ranked gene by investigating its neighborhood and the

Downloaded from nar.oxfordjournals.org at KU Leuven on May 23, 2011

neighboring genes' expression levels. Highly influencing neighboring genes are not necessarily genes that are strongly connected with the candidate gene in the network, because our method based on a random walk approach considers both the strength of the interactions and the differential expression levels of the interactors. Therefore, a candidate gene might still be highly influenced by low interacting genes with high differential expression, leading to a high ranking score.

Furthermore, if chosen by the user, PINTA provides a comparison analysis between its own ranking and the ranking obtained by simply ordering the differential expression values decreasingly in a second table using different color schemes for an easy visualization. This comparison highlights that some highly ranked genes by PINTA have a strongly differentially expressed neighborhood, although their own differential expression level is low leading to a low ranking position in the differential expression ranking. These candidate genes would not be detected by this standard analysis as our benchmark could demonstrate (18).

All output tables and subnetworks containing the strongest contributing genes can be downloaded by the user for further use. The output tables are available as tab delimited text files that can be directly opened in Microsoft® Excel as well as in any text file reader. The graphical views of the subnetworks are available as pictures (.jpg).

## SOFTWARE DOCUMENTATION

PINTA provides an online manual. In this manual, the workflow of PINTA is explained step-by-step using an exemplary expression data set. Additional pictures and screenshots can guide the user who wants to understand the details and to fine tune the prioritization parameters.

For testing purposes only, PINTA provides an exemplary expression data file that can easily be used by checking a dedicated checkbox during the fourth step of the wizard. By doing so, users find a quick way to examine PINTA's features. The data set is publicly available (GSE10849) and represents a mouse knockout experiment of the Cav1 gene. This data was RMA (24) preprocessed and as differential expression measures we computed the test statistic derived from cyberT (25). Performing a genome-wide ranking with default setting, PINTA ranks the knockout gene Cav1 on the first position, because this gene causes the most disrupted neighborhood in expression within the network.

## IMPLEMENTATION

The gene prioritization method used by PINTA was implemented in MATLAB. To make it universally accessible, we have developed a PHP web-based interface that runs with the most common web browsers, for which Java does not have to be installed. PINTA is freely accessible and there is no login requirement.

## CONCLUSION

We have developed the freely accessible web resource PINTA designed for the prioritization of candidate genes based on the differential expression of their neighborhood in a genome-wide protein–protein interaction network. PINTA is dedicated to the study of genetic disorders for which only little is known beforehand or when no background knowledge is assumed. PINTA relies on the presence of disease specific expression data, which makes it particularly attractive to study genetic conditions for which such expression data can easily be collected. PINTA propagates the expression data over the network using several random walk strategies. This allows the detection of a strong signal for a candidate gene even if its own differential expression value is small. PINTA is available for some prominent model organisms beside human and various array platforms are supported. PINTA ranks the candidate genes based on the assumption that strong candidate genes tend to be surrounded by many differentially expressed neighboring genes. A benchmarked on 40 mouse knockout experiments has shown that PINTA outperforms traditional approaches.

Downloaded from nar.oxfordjournals.org at KU Leuven on May 23, 2011

## FUNDING

The Research of D.N., L.C.T. and Y.M. is supported by Research Council KUL (ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 *SymBioSys* en KUL PFV/10/016 *SymBioSys*, START 1, several PhD/postdoc & fellow grants); Flemish Government [FWO: PhD/postdoc grants, projects, G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR), IWT: PhD grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3, FOD:Cancer plans, IBBT]; Belgian Federal Science Policy Office [IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: From Genomes to Networks, 2007–2011)]; EU-RTD (ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED); FCT PhD grant SFRH/BD/36586/2007 (to J.P.G.); FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and NEUROCLINOMICS (PTDC/EIA-EIA/111239/2009) (to S.C.M., in part).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Tranchevent,L.C., Capdevila,F.B., Nitsch,D., Moor,B.D., Causmaecker,P.D. and Moreau,Y. (2010) A guide to web tools to prioritize candidate genes. *Brief. Bioinform.*, doi:10.1093/bib/bbq007.
2. Hutz,J.E., Kraja,A.T., Mcleod,H.L. and Province,M.A. (2008) CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet. Epidemiol.*, **32**, 779–790.
3. Tranchevent,L.-C., Barriot,R., Yu,S., Van Vooren,S., Van Loo,P., Coessens,B., De Moor,B., Aerts,S. and Moreau,Y. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, **36**, W377–W384.

4. Chen,J., Xu,H., Aronow,B.J. and Jegga,A.G. (2007) Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, **8**, 392.
5. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 155.
6. Seelow,D., Schwarz,J.M. and Schuelke,M. (2008) GeneDistiller - distilling candidate genes from linkage intervals. *PLoS ONE*, **3**, e3874.
7. Yoshida,Y., Makita,Y., Heida,N., Asano,S., Matsushima,A., Ishii,M., Mochizuki,Y., Masuya,H., Wakana,S., Kobayashi,N. *et al.* (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.*, **37**, W147–W152.
8. Köhler,S., Bauer,S., Horn,D. and Robinson,P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
9. Franke,L., Bakel,H., Fokkens,L., de Jong,E., Egmont-Petersen,M. and Wijmenga,C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.
10. Radivojac,P., Peng,K., Clark,W.T., Peters,B.J., Mohan,A., Boyle,S.M. and Mooney,S.D. (2008) An integrated approach to inferring gene-disease associations in humans. *Proteins*, **72**, 1030–1037.
11. Hristovski,D., Peterlin,B., Mitchell,J.A. and Humphrey,S.M. (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inf.*, **74**, 289–298.
12. Yu,W., Wulf,A., Liu,T., Khoury,M.J. and Gwinn,M. (2008) Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics*, **9**, 528.
13. Cheng,D., Knox,C., Young,N., Stothard,P., Damaraju,S. and Wishart,D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36**, W399–W405.
14. Driel,M.A.van, Bruggeman,J., Vriend,G., Brunner,H.G. and Leunissen,J.A.M. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.: EJHG*, **14**, 535–542.
15. Van Vooren,S., Thienpont,B., Menten,B., Speleman,F., De Moor,B., Vermeesch,J. and Moreau,Y. (2007) Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res.*, **35**, 2533–2543.
16. Xiong,Q., Qiu,Y. and Gu,W. (2008) PGMapper: a web-based tool linking phenotype to genes. *Bioinformatics*, **24**, 1011–1013.
17. Nitsch,D., Tranchevent,L.-C., Thienpont,B., Thorrez,L., Van Esch,H., Devriendt,K. and Moreau,Y. (2009) Network analysis of differential expression for the identification of disease-causing genes. *PLoS ONE*, **4**, doi:10.1371/journal.pone.0005526.
18. Nitsch,D., Gonçalves,J.P., Ojeda,F., Moor,B.D. and Moreau,Y. (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, **11**, 460.
19. White,S. and Smyth,P. (2003) Algorithms for estimating relative importance in networks. *Proceedings of SIGKDD*, pp. 266–275.
20. Lovász,S. (1996) Random walks on graphs: a survey. In Milos,D., Sos,V.T. and Szony,T. (eds), *Combinatorics, Paul Erdős is Eighty*. Budapest, Hungary: János Bolyai Mathematical Society, pp. 353–398.
21. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
22. Brown,K.R. and Jurisica,I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.
23. Saad,Y. (1992) Analysis of some Krylov subspace approximations to the matrix exponential operator. *SINUM*, **29**, 209–228.
24. Irizarry,R.A., Hobbs,B., Colin,F., Beazer-Barclay,Y.D., Antonellis,K., Scherf,U. and Speed,T.P. (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
25. Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.


Downloaded from nar.oxfordjournals.org at KU Leuven on May 23, 2011

# Chapter 7

# Discussion

In this work we have first introduced gene prioritization by comparing several online accessible gene prioritization tools and by assessing their performance in real applications. Then we have introduced a novel approach for gene prioritization by replacing prior knowledge about the biological process with experimental data on differential gene expression between affected and healthy individuals. We assess a candidate gene by considering the level of differential expression in its neighborhood under the assumption that strong candidates will tend to be surrounded by differentially expressed neighbors. In *Study 3* we have presented the first implementation of this concept, followed up by an algorithmic extension and optimization in *Study 4*. The web server presented in *Study 5* enables researchers to make use of our work online.

Here, we discuss our work presented in this thesis, starting with our review about gene prioritization tools and their benchmark written in *Study 1* and *Study 2* (Section 7.1). Then we discuss the purpose and the power of our new gene prioritization approach presented in *Study 3* and *Study 4* (Section 7.2), followed by case studies and a benchmark that we set up to assess our approach's performance (Section 7.3). After that, we discuss the limitations, from which our approach still suffers, such as the quality and coverage of the underlying network (Section 7.4), the relevance of the expression data under study (Section 7.5), the nature of disease under study (Section 7.6), or the choice of an appropriate algorithm for measuring the differentially expressed neighborhood of a disease causing gene (Sections 7.7 and 7.8). After discussing these issues, they will be addressed by potential solutions in future work in Chapter 8.

## 7.1 Gene prioritization tools and their critical assessment

We have reviewed a collection of gene prioritization tools that are web-based and freely accessible and that have been developed in the past few years. In *Study 1* we have written a review to clarify gene prioritization to the biologist or clinician through a guide of 19 gene prioritization tools. This taxonomy has been done according to different characteristics of the tools, including the type of input, data sources used during the process of prioritization and the desired output. This study helps biologists or clinicians find a way through the maze of tools that have been developed throughout the last few years and to guide them in selecting the most convenient tool for their own analysis.

Since this field is still emerging, new tools will be developed in the future. To keep up with this running process, we have developed a web site, namely Gene Prioritization Portal (http://www.esat.kuleuven.be/gpp), that contains updated information of all the tools presented in this study, but also from additional tools that will be developed in the future.

To follow up this work, we have set up a benchmark in *Study 2* to critically assess the performance of some of the gene prioritization tools from *Study 1*. The benchmark consists of 42 (recently discovered) disease-gene associations aiming at comparing seven different gene prioritization tools (SUSPECT [20], ToppGene [34], GeneDistiller [15], GeneWanderer [85], Posmed [100], CANDID [39], and Endeavour [50]) under the same conditions and providing that the recently published disease-gene associations would be unknown. To reach this goal, we have run the experiment on our selection of tools very close to the publication time of the disease-gene association. By doing this, we tried to assure that the data sources or databases behind the gene prioritization tools were not updated by the novel disease-gene association. Unfortunately, we had to exclude gene prioritization tools, such as PolySearch [12], because it uses up-to-date data retrieved from Pubmed. However, we could not completely guarantee that none of the gene prioritization tools would be updated while we were performing the experiments, but we could at least minimize this risk factor.

A limitation that occurred in line with our benchmark was the fact that among the disease-gene associations we could identify weak (not confirmed) associations beside strong (confirmed) associations. An example for a strong and confirmed disease-gene association is the Ehlers-Danlos syndrome, caused by loss-of-function mutations of the CHST14 gene. Miyake and colleagues [68] have investigated a new type of the Ehlers-Danlos syndrome, similar to kyphoscoliosis but without lysyl hydroxylase deficiency. They have identified a homozygous CHST14 (carbohydrate sulfotransferase 14) mutation in the two familial cases and compound heterozygous mutations in four sporadic cases. CHST14 encodes D4ST1, an important membrane

protein, which is absent in patients, and the loss of the decorin DS proteoglycan due to CHST14 mutations may preclude proper collagen bundle formation or maintenance of collagen bundles, which plays an important role in the Ehlers-Danlos syndrome.

## 7.2 The purpose of expression based *a priori* gene prioritization

Genetic studies, including cytogenetic, linkage, and association studies, can identify chromosomal regions associated with a disease or phenotype of interest. However, biologists are often confronted with long lists of candidate genes among which they need to select a limited number of candidates for further validation. In the past years, computational methods for prioritizing candidate genes have been proposed. They usually rank candidates by matching their information across multiple data sources against a profile derived from a set of genes, pathways, or biological processes already known to be involved in the phenotype. Most of these available prioritization methods (Endeavour, Prioritizer, ToppGene, etc) have in common that they require knowledge about the disease to identify putative disease genes, for example in the form of a set of genes, pathways, or gene ontology categories known to be implicated in the disease. They then rank candidate genes through guilt-by-association methods across a variety of data sources. But when nothing or only little is known about a disease, these methods will be inapplicable or ineffective. We seek to overcome this limitation by replacing knowledge about the biological process by experimental data on differential gene expression between affected and healthy individuals. Considering the problem from the perspective of a gene or protein network, we assess a candidate gene by considering the level of differential expression in its neighborhood under the assumption that strong candidates will tend to be surrounded by differentially expressed neighbors.

A standard genetic procedure used to analyze candidate genes is to check the expression level of a candidate gene in patient derived material against wild type (typically in fibroblast or immortalized lymphoblastoid cell lines). Candidates for which a significant difference is observed between the two groups are considered promising. However, for many genetic diseases (such as diseases arising from point mutations in coding regions), there is no guarantee that the expression level of the disease gene itself is affected (although this is possible through feedback effects). Rather, genes that are downstream of the disease gene are those whose expression will be affected. Instead of considering genes in isolation, we consider the differential expression data now at the level of a gene or protein network. While it is extremely useful to incrementally add disease genes to phenotypes whose molecular basis is reasonably well characterized, there is strong demand from geneticists for methods that could help in the more difficult case of disorders for which the molecular basis

is not yet elucidated. Truly effective prioritization methods for this case had so far been lacking.

## 7.3 Case studies & benchmark study

In *Study 3* we have illustrated our approach on four monogenic diseases using publically available expression data from GEO: Fragile X syndrome (FXS, GSE7316) caused by mutation of FMR1 (fragile X mental retardation 1), Marfan syndrome (MFS, GDS2960) caused by mutation of FBN1 (fibrillin 1), Cystic fibrosis (CF, GDS2143) caused by mutation of CFTR, and Becker muscular dystrophy (BMD, GDS2855) caused by mutation of DMD. We could successfully rank the disease genes in high positions compared to other genes from the same chromosomal regions. Then, we have applied the same approach on a polygenic disorder, Stein-Leventhal syndrome or polycystic ovary syndrome (PCOS) using a publically available expression data set (GEO accession number GDS2084). For PCOS no disease gene was known, but we could confirm two genes that are likely to be involved in this disease: fibrillin 3 (FBN3, for which a susceptibility allele has been identified) [61] and follistatin (FST, for which association with a Stein-Leventhal syndrome related metabolic phenotype has been shown) [60]. Furthermore, we have suggested a new candidate gene (DEAD box 4) having potential causal effect in PCOS because DDX4 was ranked first among all genes from 5q11.2 and because DDX4 is expressed in the murine ovary and involved in the growth of oocytes and follicles of mice. Recently, we have repeated the gene prioritization for PCOS using the best performing algorithm from *Study 4*, the Heat Kernel Diffusion Ranking. By doing this, we could confirm our findings from *Study 3*: FST was ranked at position 4 out of 29 genes with a p-value of 0.06, and at position 464 in a genome-wide ranking (out of >17,000 genes). DDX4 was again ranked at position 1 with a significant p-value (0.006), see Table 7.1. Although FST has been shown not to be a disease causing gene for PCOS, but against the background of its known properties (such as its metabolic role), we expect it to rank high since it is likely to be involved in PCOS.

PINTA (*Study 5*) enables the user to further explore the genes with high ranking scores. Figure 7.1 illustrates the contribution of neighboring genes of FST contributing to its high ranking score, highlighting that most contribution is coming from several highly differentially expressed genes in a close neighborhood, such as INHBA, GPRC5A, and BMPR1B. This fact demonstrates the power of a random walk approach propagating the expression signal through the network. Due to this approach, we could rank FST highly, although its own differential expression level was low (-0.11).

In *Study 4* we have compared the performance of our random walk based method with a standard procedure in genetics that ranks candidate genes based solely on

| Gene | DE (log ratio) | score | p-value |
|---|---|---|---|
| DDX4 | -0.17 | 0.2 | 0.006 |
| PDE4D | -0.19 | 0.14 | 0.044 |
| ITGA2 | -0.06 | 0.13 | 0.048 |
| **FST** | **-0.11** | **0.14** | **0.066** |
| MAP3K1 | -0.16 | 0.15 | 0.102 |
| ARL15 | -0.14 | 0.08 | 0.116 |
| CCNO | -0.22 | 0.14 | 0.122 |
| PLK2 | -0.14 | 0.1 | 0.148 |
| ISL1 | 0.1 | 0.11 | 0.154 |
| GZMK | 0.04 | 0.06 | 0.214 |
| ESM1 | 0.13 | 0.09 | 0.216 |
| IL6ST | 0.03 | 0.11 | 0.318 |
| IL31RA | 0 | 0.03 | 0.334 |
| GZMA | 0.02 | 0.06 | 0.474 |
| ACTBL2 | 0 | 0.03 | 0.532 |
| SKIV2L2 | -0.01 | 0.03 | 0.554 |
| MOCS2 | -0.01 | 0.03 | 0.558 |
| DHX29 | 0.02 | 0.02 | 0.62 |
| PPAP2A | 0.02 | 0.05 | 0.774 |
| ITGA1 | 0 | 0.07 | 0.958 |
| CDC20B | 0 | 0 | 1 |
| GAPT | 0 | 0 | 1 |
| GPX8 | 0 | 0 | 1 |
| GPBP1 | 0 | 0 | 1 |
| C5ORF35 | 0 | 0 | 1 |
| SNAG1 | 0 | 0 | 1 |
| RAB3C | 0 | 0.01 | 1 |
| NDUFS4 | 0 | 0.02 | 1 |
| HSPB3 | -0.07 | 0.04 | 1 |

Table 7.1: Gene prioritization of 29 genes coming from the chromosomal region 5q11.2 for the polycystic ovary syndrome (PCOS) based on the Heat Kernel Diffusion Ranking implemented in *Study 4* and *Study 5* and the STRING network (v.8.2) and public expression data (GDS2084).

their differential expression levels. To analyze whether we can rank FST highly based on its differential expression value only, we have ranked the same 29 genes as before by their expression. Since we could already see in Figure 7.1 that FST would not be highly differentially expressed compared to other genes, we did not expect FST to rank highly. In fact, FST was ranked at position 22 out of 29 genes with a low differential expression level of -0.11, whereas the highest differential expression signal on this chromosomal region was coming from the CCNO gene (-0.22). The

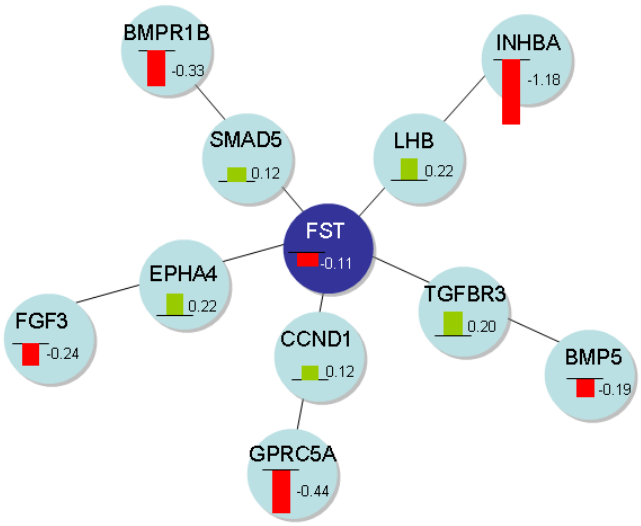



Figure 7.1: FST subnetwork containing the strongest contributor genes for the high ranking score of FST derived from PINTA, based on the STRING network v.8.2 and public expression data (GDS2084).

CCNO gene was ranked by PINTA at position 7 only, although this gene is more highly differentially expressed than the FST gene. This demonstrates the strength of our method: it can rank a lowly differentially expressed gene on a top position as long as its interaction partners in the network are highly differentially expressed. Due to this fact, CCNO was ranked on a lower position than FST because of the weaker expression signal in its neighborhood (see Figure 7.2).

The first version of our method was applied on four distinct monogenic diseases to estimate its performance in *Study 3*. Needing more validation, we have set up a benchmark in *Study 4* containing 40 knockout mouse models for assessing the performance of different algorithms implemented in our method against results obtained using the standard procedure ranking genes solely on their expression. Here, results showed that our strategies could outperform this standard procedure for all implemented algorithms. This finding allowed us to state that our method can identify promising candidate genes using network-based machine learning approaches even if no knowledge is available about the disease or phenotype.

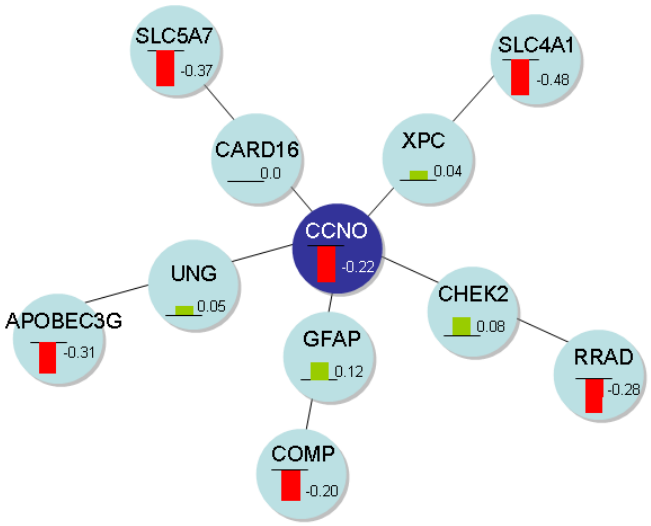



Figure 7.2: CCNO subnetwork containing the strongest contributor genes for its ranking score derived from PINTA, based on the STRING network v.8.2 and public expression data (GDS2084).

## 7.4 The quality and coverage of the underlying network

The quality and coverage of the underlying network is a strict bottleneck in our work presented in *Study 3* and *Study 4*. To evaluate to what extent the coverage of the interaction network influences our method, we have applied two distinct PPI networks of different coverage in *Study 4*, namely BioGRID [8] and OHPHID [46], and the STRING network [53] as a good-coverage functional association network. Both, the STRING and the OHPID network, integrate the BioGRID database among others in their network, and we could expect them to perform better than BioGRID alone due to the extra information. The results of this study have clearly revealed that the BioGRID network was too sparse for our application. The mouse version was composed of only 1,417 proteins and 2,026 interactions, in comparison with the OHPID network (10,867 proteins and 79,088 interactions) and the STRING network (24,442 proteins and 1,405,375 associations) being a sparse network. Because of the best network coverage, our method has performed best in this study using the STRING network. At first sight it might

not be appropriate to compare two PPI networks with a functional association network because the latter includes predicted associations derived from literature as well as physical interactions deduced from distinct PPI databases, including BioGRID. But in this study we were searching for a high coverage and high quality network that would perform best in our application regardless of the nature of its interactions or associations. For this reason, we could recommend the use of the STRING network as a good-coverage functional association network for our method. Nevertheless continuing improvements in the quality of protein association networks might contribute to further increase the effectiveness of our method.

Here, the quality of the network was only assessed by the network coverage. However, STRING is mainly based on literature entailing errors (caused by incorrect gene annotation, unreliable functional annotation, etc.) that may both cause false negative (missing associations between proteins) as well as false positive associations (incorrect associations between proteins). Due to manual curation of the PPI data they use, BioGRID and OHPID are PPI networks of high quality. Nevertheless, the performance of our method using the STRING network was better than using the other two networks. Therefore, we can assume that the coverage of the network can be stated as the main bottleneck of our method, whereas the quality remains an important issue but might be increased by the coverage and the continuing improvements in its quality.

## 7.5 The quality and the relevance of the expression data

Another main bottleneck of our method is the quality and the relevance of the underlying expression data. Poorly collected samples (patients not actually sharing the same molecular phenotype) or disease heterogeneity (multiple genes or pathways leading to similar phenotypes) can obscure the expression pattern and will make picking up a meaningful signal more difficult. Similarly, lack of access to the most relevant cell types (biopsies cannot be performed arbitrarily on patients) can be a limiting factor for the method. Our method is based on the idea that the pattern of differential expression is concentrated on the network to highlight the affected genes and its surrounding interaction partners. We have used mouse knockout data sets in *Study 4* guaranteeing the use of disease relevant tissue which is an important issue in identifying disease related genes, because disease genes might not be expressed in tissues that are not causing disease or phenotype [33]. Applied to human diseases that could mean to collect samples from disease relevant tissues (*e.g.,* muscle biopsy from a patient with a muscle disease). However, this procedure is not applicable for many diseases in which tissues are affected from which no samples can be collected *in vivo* (*e.g.,* brain biopsy from a patient with mental

retardation). Therefore, the relevance of the expression data is still a challenge that we have addressed in Section 8.1 suggesting feasible solutions.

## 7.6 Monogenic versus polygenic and complex diseases

The aim of our method is tailored to monogenic diseases (inherited and controlled by a single pair of genes), assuming that the gene causing the largest disruption in expression in the network is likely to be the disease causing gene. Our benchmark in *Study 4* was built up from distinct mouse knockout data sets, guaranteeing that the phenotype is caused by a single gene mutation and expecting that the disruption in expression can be clearly identified. This can be intuitively transfered to monogenic diseases, such as Fragile X syndrome, Becker muscular dystrophy, or Cystic fibrosis as analyzed in *Study 3*. Although it would be preferable to detect the disease gene, the ability of the method to return a negative result is an asset of the method. On the one hand, we expect that our method might be more broadly applicable, though, as demonstrated on our case study on the polygenic disorder polycystic ovary syndrome (PCOS), where we could detect the FBN3 gene and the FST gene as related to the disease. On the other hand, we have also applied our method on distinct polygenic and complex diseases, such as breast cancer and diabetes, where we could not detect disease related genes. For the latter cases, one or several disrupted expression modules were found but none of them was surrounding a disease related gene. The same observation could be made for a few single gene disorders, such as the Rett syndrome or the Scott syndrome. This could be explained by the fact that even a single-gene disorder can result in the disruption of several pathways (downstream of the original pathway), making therefore the signal we are looking for more difficult to detect. This reflects that the current method is not suitable for all single gene disorders, nor for complex diseases.

## 7.7 Measuring differential expressed neighborhoods of disease genes

In *Study 3*, we have introduced our method by assessing the probability of a gene to be involved in a disease under study by measuring its differential expressed neighborhood in the network. Our algorithm was based on a notion of soft neighborhood where each gene is given a contributing weight, which decreases with the distance from the candidate gene on the protein network. To account for multiple paths between genes, we have defined a notion of distance using the Laplacian exponential diffusion kernel. We have scored candidates by aggregating

the differential expression of neighbors weighted as a function of distance. To study the neighborhood of candidate genes in the distance network, we had to bound it to a limited size because the network was densely connected, and thus the neighborhood of a candidate gene consisted of almost all genes in the network. This issue was addressed by considering the neighboring genes with the smallest distances to the candidate (any gene further than a certain threshold was not considered). We determined the size of the neighborhood by analyzing the observed signals obtained by applying a test statistic. However, we obtained inconsistencies in the size of the neighborhood across several diseases under study, leading to the need for an optimized approach that is presented in *Study 4*.

Here, we started to overcome the limitation of *Study 3* by developing new algorithms. Instead of working on a distance network and considering the closest genes in its neighborhood, we implemented random walk based algorithms, propagating the expression signal through the network. However, we still had to determine the neighborhood size indirectly, by defining the number of steps executed in the random walk. But we could demonstrate that a reduced number of iterations is usually sufficient for ranking purposes using our best performing algorithm, the Heat Kernel Diffusion Ranking, in agreement with another study [6]. However, the random walk approach came along with a systematic bias using a biological network favoring highly connected genes leading to sensitivity to skewed degree distributions. This issue could be addressed through a randomization procedure that we have not yet implemented and should be done in the future.

## 7.8 PINTA - a web server for gene prioritization

To finish up our work, we have developed PINTA (www.esat.kuleuven.be/pinta), a web resource for the prioritization of candidate genes based on the differential expression of their neighborhood in a genome-wide protein-protein interaction network. In *Study 5*, we have implemented a web server using the best performing ranking algorithm and network obtained from *Study 4* as default setting. The user can assess our gene prioritization method by using this freely accessible and user-friendly web tool. PINTA supports both candidate gene prioritization (starting from a user defined set of candidate genes) as well as genome-wide gene prioritization and is available for five species (human, mouse, rat, worm, and yeast). As input data, PINTA only requires disease specific expression data, whereas various platforms (*e.g.*, Affymetrix) are supported. As a result, PINTA computes a gene ranking and presents the results as a table that can easily be browsed and downloaded by the user.

PINTA was developed to make our method freely accessible for every user and uses by default the best performing settings (the Heat Kernel approach and the STRING network) based on our findings from the benchmark, which is convenient

for nonexpert users. Advanced users can fine tune the parameter to fit their needs and make the best of their data by choosing between five different prioritization algorithms (the Heat Kernel and Arnoldi Diffusion approach from *Study 3* plus a simple random walk implementation, the HITS with priors and the k-step Markov approach).

As we have discussed in Section 7.6, our method is not suitable for all diseases. If the user runs PINTA with the default setting and is not convinced by the ranking result, she can easily redo the gene prioritization by choosing a different algorithm than the default. In our benchmark we could observe that for some mouse knockout experiments the Heat Kernel approach could not detect the sought gene but one of the others could. Therefore, the user could obtain a different (potentially a better) result applying a different algorithm.

# Chapter 8

# Perspective

As discussed in Chapter 7, the main bottlenecks of our work are the quality and coverage of the underlying network and the relevance of the expression data. To address the latter issue, we propose to use tissue-specific expression data based on either a publically available global gene expression map or based on induced pluripotent stem (iPS) cells. The global gene expression map could be applied easily, although it is available for a few diseases only, whereas using tissue-specific expression data derived from iPS cells is laborious, time- and money-consuming, but could be derived for each tissue and disease under study. We propose to first start with the global map of expression data to estimate the gain of information we may retain using tissue-specific expression data compared to non-tissue-specific expression data. In a second step, studies could be set up for certain diseases of interest using expression data from iPS cells for further investigation (Section 8.1).

Integrating several data sources could be suitable to remove the number of false positives and false negatives and to improve the quality of the ranking results. Here, we propose to extend our approach by using an additional layer with extra information (*e.g.,* gene annotation), to ascertain similarities between genes (Section 8.2).

Another concept to study the network-based change of expression could be the analysis of time series gene expression data that could be useful in studying drug response or disease development. Instead of considering expression data from only one time point, the expression data from various time points would be assessed to analyze which gene's neighborhood changes its expression over time or under treatment (Section 8.3).

To use our method for discovering novel disease genes, we suggest to continue and extend existing and novel collaborations with clinicians. The final goal would be

not only to better understand the disease under study, but also to elaborate on drug development for personalized medicine (Section 8.4).

Beside the perspective projects, two ongoing projects are described, such as setting up a pipeline through gene prioritization tools as an extension of *Study 1* and *Study 2*, for identifying novel disease genes. Here, we combine different tools starting at the genome level, determining a set of candidate genes and heading towards a short ranking list of promising candidate genes that are suitable for further investigation (Section 8.6). Additionally, we study the tissue-specificity of protein-complexes based on expression data for detecting complexes that are co-expressed in normal tissues in which they cause disease phenotype or to discover pleiotropic effects (Section 8.5).

## 8.1 Using tissue-specific expression data for quality improvement

As discussed in Section 7.5, a major limitation of our work using disease-specific expression data for network analysis is the informativeness and relevance of the expression data under study. For many disease-specific expression data sets EBV cell lines are used because disease-specific tissues are not always available: Nishimura and colleagues [98] analyzed lymphoblastoid cells from patients with autism, and Yao and colleagues [102] analyzed skin fibroblast cultures from patients with Marfan syndrome, whereas Bakay and colleagues [54] could analyze disease specific muscle tissue in patients with various muscle diseases. However, if expression data comes from a non disease-specific tissue, it cannot be guaranteed that relevant differences can be detected between patients and healthy controls due to the different pathways activated in different tissues [74] [64].

Here, we propose two concepts addressing this issue. First, we suggest a concept based on a publically available global map of human gene expression [56] for estimating the gain of information we would obtain when using tissue-specific expression data compared with non-tissue-specific expression data. If the information gain is large enough to demand for further network analysis based on tissue-specific expression data, the second step could be initiated, using induced pluripotent stem (iPS) cells for deriving tissue-specific expression data for a specific disease under study for which no expression data is publically available.

### 8.1.1 A global gene expression map for tissue-specificity

Lukk and colleagues [56] have set up a publically available global map of human expression data coming from 5,372 samples, representing 128 different tissues in 4

different cell types (normal, disease, neoplasm, and cell line). Among these, they provide for 18 tissues normal status beside disease status as shown in Table 8.1, whereas for some tissues several diseases are available. However, for some diseases only few samples can be found in the data set, (*e.g.,* for the lung disease emphysema or HIV only one sample each is available). Therefore, we would carefully consider the number of samples of normal and disease tissue that are necessary before we would start computing differential expression levels. This data set could be a starting point for a new benchmark for our method assessing its performance in human by using tissue-specific expression data versus non-tissue-specific expression data (as used in *Study 3* and *Study 4* ), expecting the first to perform better due to its relevance. By doing this, we could estimate the error retrieved by using non-tissue-specific compared to tissue-specific expression data which could give us an idea of performance gain if the user provides tissue-specific expression data using PINTA. To set up this benchmark, we could use the disease data sets from [56] as described before (see Table 8.1), comparing to data sets of non-tissue-specific expression for the same diseases, but originated from different tissues than the disease tissues (*e.g.,* lymphoblastoid cells from patients with Huntington disease, see GEO accession number GSE23834 [65]).

### 8.1.2 Induced pluripotent stem cells for tissue-specificity

Induced pluripotent stems (iPS) cells have been artificially reprogrammed from human somatic cells while exhibiting the essential characteristics of embryonic stem (ES) cells and maintaining the potential of differentiating into all tissue cells [38]. Several studies have demonstrated that disease specific iPS cells can be generated that can be useful for the development of novel disease models and in drug discovery [99] [31]. For example, Dimos and colleagues [40] could generate iPS cells from patients with amyotrophic lateral sclerosis (ALS) that in turn could be differentiated into motor neurons, the cell type destroyed in this disease.

Using expression data from iPS cells derived from patients would enable us to differentiate them into the cell type in which the disease occurs, such as neurons for autistic patients. This technique would have the power that expression data from disease-specific tissue could be made available even if a tissue biopsy is (1) not practicable *in vivo* (such as a brain biopsy to get neuron cells from autistic patients or healthy controls) or (2) uncomfortable (such as a muscle biopsy). Instead, lymphoblastoid cell lines could be used for all kind of diseases that are obtained by EBV transformation of blood cells which is the easiest, most accessible source of cells [26] [49], or from skin fibroblast biopsies (as derivation of iPS cells from EBV cell lines has not been proven). This technique would improve the quality of analysis of disease-specific expression data significantly, especially for diseases in which the relevant tissue is barely or not accessible *in vivo*, such as for mental retardation. However, this technique is still in its infancy but already shows nowadays promising

| Tissue | Diseases |
|---|---|
| Adipose | Polycystic ovary syndrome |
| Atrial myocardium | Ischemia |
| Cardiac ventricle | Myocardial infarction, Atrial fibrillation |
| Blood | Hepatitis C, Urinary tract infection, Influenza, Malaria, Osteomyelitis, Trauma, Pneumonia, Abscess, Common variable immunodeficiency, X-linked agammaglobulinemia, Down syndrome, transient Myleoproliferative disorder, Meningitis, Tendon xanthomas, Atopic severe asthma, Bacteremia, Endocarditis, Heart transplant rejection, Respiratory distress, Septic arthritis, Atopic mild asthmatics, Cellulitis, HIV, Seizure |
| Bone | Hyperparathyroidism |
| Bone Marrow | Monoclonal gammopathy, Down syndrome |
| Brain | Bipolar disorder |
| Cerebellum | Huntington disease |
| Caudate nucleus | Huntington disease |
| Frontal cortex (primary) | Huntington disease |
| Frontal cortex (superior) | Huntington disease |
| Hippocampal CA1 tissue | Alzheimer disease |
| Esophagus | Barretts esophagus |
| Lung | Emphysema |
| Ovary | Polycystic ovarian syndrome |
| Quadriceps Muscle | Duchenne muscular dystrophy |
| Skeletal Muscle | Juvenile dermatomyositis, Mitochondrial disorder, Facioscapulohumeral muscular dystrophy, Emery-Dreifuss muscular dystrophy, Duchenne muscular dystrophy, Calpainopathy, Metabolic syndrome, Dermatomyositis, Congestive heart failure, Hereditary spastic paraplegia |
| Skin | Dermatitis |

Table 8.1: The global map of human expression data [56] was made available for 18 tissues with normal and disease status.

results, such as Jozefczuk and colleagues [35] recently compared hepatocytes from ES cells with hepatocytes from iPS cells and discovered that their gene expression is similar which could lead to a treatment of liver diseases in the future. Similarily, Soldner and colleagues [25] recently showed that fibroblasts from patients with Parkinson's disease can be efficiently reprogrammed and subsequently differentiated into dopaminergic neurons.

## 8.2 Data integration for reducing false positives and false negatives

To identify novel disease genes, enormous data is needed and high-throughput studies are often used to gather important information about potential disease-gene associations. Gene prioritization methods can simplify this search but they often suffer from many false positives and false negatives. To overcome this limitation, methods have been developed that integrate distinct data types to reduce the number of false positives and false negatives by eliminating inconsistencies. For example, Aerts and colleagues [81] [50] [16] developed ENDEAVOUR, a gene prioritization method that ranks candidate genes based on their similarity to genes already involved in the biological process of interest, using multiple data sources, such as ontologies and annotations, protein-protein interactions, cis-regulatory information, gene expression data sets, sequence information and text-mining data, and merging the individual rankings based on the distinct data sources into a global ranking of the candidate genes. The same applies to a number of other gene prioritization methods, such as CANDID [39], Genedistiller [15], or ToppGene [34], which could improve their performance by data integration.

Pers and colleagues [90] developed an approach, called MetaRanker, to combine heterogeneous data from genome-wide association studies, protein-protein interaction screens, disease similarity, linkage studies, and gene expression experiments into a multi-layered evidence network, which is used to prioritize the entire protein-coding part of the genome identifying a shortlist of candidate genes for complex phenotypes. From their point of view, our method consists of a two-layer network: the first layer containing the weights of protein-protein interactions, while the second including expression data. Considering the work of Pers and colleagues, a logical step in extending our work could be to set up additional layers of further data types. However, still keeping our concept of studying unknown diseases, we would avoid including any data types that resemble prior knowledge, such as literature or training genes (*i.e.,* genes known to be involved in the phenotype or disease of interest). A feasible data type could be data coming from Gene Ontology (GO) that allows us to annotate genes and their products with a limited set of attributes and that describes how gene products behave in a cellular context [11]. Since GO is structured as a directed acylic graph, we could define a measure of similarity between genes as done before with interaction data. Genes with similar or equal attributes would then obtain a large similarity or small distance. Extending our existing two-layer network by a new layer, we would add information about genes that are similar in terms of their annotation.

However, the majority of data types that have been used by other gene prioritization methods for data integration, is coming from protein-protein interaction and gene annotation on the one hand, and from text-mining, sequencing, gene-phenotype

associations or training genes on the other hand. The latter data types all require prior knowledge about the disease. For a known disease: publications, sequence data, gene-phenotype associations or training genes are available. However, genes can be known and well studied in terms of their cellular context, function or regulation, without being aware of a contingent involvement in a certain disease. Therefore, we would concentrate on integrating those data types that only rely on this kind of information.

The computational implementation of this multi-layered network could be done using kernel methods for data fusion resulting in a separate kernel matrix for each layer and fusing the kernel matrices by uniform linear combination as presented in an extension of ENDEAVOUR [16], a Bayesian network for data integration [71] or a multi-layered Bayesian network [86]. A straightforward concept for setting up the multi-layered network could be to compute ranking scores for each layer separately and combining these scores into a final and global score. Based on our Heat Kernel algorithm, the random walk on the network would be performed on each layer independently (*e.g.,* expression data and gene annotation), resulting in two distinct ranking lists that could be combined into a global score, similarly as in [81] [90].

## 8.3 Network analysis of time series gene expression data for studying drug response

In static expression experiments, a snapshot of the expression of genes in different samples is measured, while in time series expression experiments, a temporal process is measured. At any time point, mRNA is transcribed and proteins are generated, dependent on functionality and environmental condition. Transcription factors (TF) regulate the expression of other genes by initiating or repressing transcription and are influenced by environmental changes (*e.g.,* drug treatment). Taking a snapshot of the expression profile after a new environmental condition has been established, new genes can be observed that are specifically expressed under this condition. However, a time series of gene expression can reveal the complete set of genes that are expressed under these conditions [101].

Another promising experimental direction for focusing the expression patterns towards the disease-causing gene would be to use more sophisticated factorial disease for the microarray experiment. If we can identify a stimulus affecting disease in individuals, we could attempt to perform a time series gene expression profiling experiment where patients receive a treatment triggering the drug treated cascades. The differential expression response will tend to be more tightly focused towards those genes that are essential for the drug response and might reveal new insights into drug development. The expression data set from Turton and colleagues

[70] could serve as an example, studying the effects of long-term treatment with tamoxifen on endometrial gene expression in women (GEO accession number GSE3626) aiming at examining changes in gene expression of the endometrium exposed to long-term tamoxifen treatment in comparison to age matched controls. Another example could be the expression data set from Zhao and colleagues [41], as they study growth hormone regulation of lipid metabolism in adipose tissue by microarray analysis in hypopituitary men before and after one month of GH treatment (ArrayExpress accession number E-MTAB-276). Both experiments study treated and untreated patients compared to healthy controls to analyze the effect of a drug under study. If we can identify a disrupted module in expression in the network that is differentially affected in treated and untreated patients and controls, we could suggest promising triggers that could be involved in disease and drug response.

## 8.4 Disease gene discovery for personalized medicine

As defined by the Personalized Medicine Coalition [10], *personalized medicine* is the use of new methods of molecular analysis to better manage a patient's disease or predisposition towards a disease. The goal is to choose optimal medical treatment tailored to a patient's genetic and environmental profile. In the future, novel diagnostic and prognostic tools will increase the accuracy of predicting a patient's outcome after drug therapy, and the identification of novel disease genes will tend to speed up and improve drug development. In cancer research, promising findings have been discovered by using specific agents directed at unique molecular features of cancer cells and and immunotherapy approaches modulating the tumour immune response, both techniques producing greater effectiveness with less toxicity [80]. In other cases, patients may not respond favorably to drug therapy or suffer severe adverse drug effects [96]. This need for supporting the development of patient-specific drugs that have a positive therapeutic effect on the disease outcome with minimal side effects favors the identification of novel biomarkers or disease genes.

A latter task for our work is aiming for the identification of novel disease genes. Since our method is predictive and we cannot reach 100% accurracy, we can only suggest novel disease genes that have to be followed up under further biological investigation, such as genome-wide association studies or sequencing. However, if a promising candidate gene derived from PINTA ends up as a real disease gene, we would have successfully completed our search for a disease causing gene for a certain disease.

**Polycystic ovary syndrome**
In *Study 3* we have suggested a novel candidate gene (DDX4) to have causal effect in the polycystic ovary syndrome (PCOS). The expression of DDX4 is restricted to the ovary and sperm and in the ovary of a mouse model, and it is specifically

expressed in primordial and growing oocytes and follicles. Until recently, this gene has not yet been reported as a candidate for PCOS and does not belong to a family or a pathway that has been previously considered in PCOS.

We have contacted the group of Prof. Bronwyn Stuckey and Prof. Scott Wilson from Queensland University of Technology (Australia) for a validation of DDX4 in a case-control association study under the hypothesis that a common genetic variation in the DDX4 gene region (chr5:55055448-55159447, NCBI Build 36) is associated with PCOS and alterations in key PCOS phenotypes. The aim of this study was to characterize the common genetic variation in the DDX4 gene region using haplotype tagging SNPs in well characterized PCOS subjects and controls and test for association with disease, and for a role in the regulation of androgen and carbohydrate metabolism. For the study, a cohort of 188 Caucasian women was recruited from the database at The Keogh Institute with a similar number of control women available that was recruited from the general population of Perth, Australia. The group could find a weak evidence of an association between SNPs in the DDX4 region (notably rs2305123), and Free Androgen Index, a key diagnostic metric for POCS. They stated that replication would be required to confirm the result in a Caucasian population, and further studies would be needed in subjects of different ancestry to determine SNP frequencies and the relevance of the observed effect in other ethnic groups.

For the replication of an association study we have contacted the group of Prof. Margrit Urbanek from Northwestern University (USA) who could not detect a significant evidence for association between variants in DDX4 and PCOS in a cohort of 867 cases and 965 controls. Cases and controls were all of European ancestry.

Nevertheless, we are still convinced by the fact that further replication has to be done in the future, although the first replication did not find evidence for an association, but the fact the first study could find weak evidence should not be neglected.

**Endometriosis**
In a recent study we have performed a global screening of differentially expressed genes in eutopic endometrium from patients with and without endometriosis during early luteal and menstrual phases of the cycle using mRNA microarray and proteomic technology. Since no differentially expressed genes were found in disease versus control, we were up to applying our method on this data set because it can detect a strong signal for a candidate gene even if its own differential expression value is small.

So far, we could not suggest novel candidate genes to be involved in endometriosis compared to healthy controls in a genome-wide ranking. The first step would be to reduce the number of genes as input by defining either a set of promising candidate genes or a chromosomal region. Then we would not only prioritize the genes for

distinct cell cycles in patients with endometriosis, but also in healthy individuals to be able to suggest novel and promising candidate genes that are distinct in both groups.

## 8.5 Tissue-specificity of protein-complexes based on expression

Defects in human genes leading to a phenotype can be highly tissue-specific. Therefore, lots of efforts in research have been made on understanding tissue-specificity of human diseases, focusing on analyzing single disease genes in affected tissues. Although it has been shown that disease genes generally tend to be expressed in a limited number of tissues, it is still unclear in many cases how the tissue-specific expression patterns of disease genes correlate with their pathological manifestations. Lage and colleagues [43] have developed a strategy for systematically correlating pathological manifestations of diseases with expression patterns of implicated genes and protein complexes across many human tissues. They detected a significant tendency for disease genes and complexes to be over-expressed in normal tissues where defects caused pathology. To do so, they have used the GNF tissue atlas [4] delivering tissue-specific expression data for 73 non-disease tissues and 1054 diseases derived from OMIM resulting in a disease tissue association matrix.

Taylor and colleagues [33] have examined the dynamic structure of the interactome to determine whether changes in the organization of the interactome can be used to predict breast cancer patient outcome. To do so, they have analyzed hub proteins in the interactome and identified intermodular hub proteins that are co-expressed with their interacting partners in a tissue-restricted manner and intramodular hub proteins that are co-expressed with their interacting partners in all or most tissues. They have used the GNF tissue atlas [4] delivering tissue-specific expression data for 73 non-disease tissues, and quantified the extent to which a hub and its interacting partners were co-expressed in the same tissues. They have used the average Pearson correlation coefficient (PCC) of coexpression of a hub protein and its partners to identify whether interactions are context specific (that is, interacting proteins are not always coexpressed) or constitutive (that is, interacting proteins are always coexpressed).

**Coordinated expression in protein complexes**
In an ongoing and collaborative study with Tune H. Pers and Prof. Søren Brunak from the Center for Biological Sequence Analysis at DTU in Copenhagen, Denmark, we are aiming at identifying the tissue-specificity of a broad set of protein complexes defined in the InWeb network (as defined in [42]) based on human gene expression data [56]. We want to analyze the tissue-specificity of protein complexes and the

properties of complexes specific to a given tissue for detecting complexes that are coexpressed in normal tissues in which they cause disease phenotype and to discover potential pleiotropic effects on protein complex level.

As a protein network on a protein complex level, we use the InWeb v4.1 network as defined in [42] that consists of approximately 20,000 proteins with 8,000 protein complexes among which 3,500 contain at least five interactor proteins besides the hub protein. For our study we consider only the latter, as these proteins are in the top 15% of the degree distribution of the network. As expression data we use a global map of human gene expression [56] coming from 5,372 samples, representing 128 different tissues in four different cell types (normal, disease, neoplasm, and cell line). Among these, 64 normal tissues are available with up to 416 samples per tissue. Here, we consider all normal tissues from this expression data set with at least five replicates, leading to 39 different tissues (see Table 8.2).

| Category | Tissues |
|---|---|
| Blood | Blood, Leukocyte, Cord blood, Fetal blood |
| Bone Marrow | Bone Marrow |
| Bone | Bone |
| Cardio vascular system (Heart) | Atrial myocardium, Cardiac ventricle |
| CNS(Brain) | Brain, Caudate nucleus, Cerebellum, Hypothalamus, Frontal cortex (primary), Frontal cortex (superior), Hippocampal CA1 tissue |
| Digestive tract | Small intestine, Esophagus |
| Endocrine Glands | Thyroid |
| Eye | Eye |
| Femal reproductive system | Myometrium, Ovary |
| Immune System | Thymus, Lymph node, Tonsil, Palatine tonsil |
| Lipocytes | Adipose tissue |
| Male reproductive system | Prostate |
| Muscle | Skeletal muscle, Smooth muscle, Quadriceps muscle |
| Placenta | Placenta basal plate |
| Respiratory system (Lung) | Lung, Lung transplant, Bronchial epithelia, Bronchial epithelia (current smoker), Bronchial epithelia(former smoker) |
| Skin | Skin |
| Urinary tract (Kidney & Bladder) | Kidney, Mesagnium |

Table 8.2: The global map of human expression data [56] was made available for 39 normal tissues with at least five replicates.

So far, we have computed the coexpression for all 3,500 protein complexes over all samples in each tissue. For each hub protein the average Pearson Correlation Coefficient of coexpression was assessed using a similar algorithm as described in [33]. To compute the correlation between a hub protein and all its interactor proteins in a complex, the coexpression for each hub-interactor pair is computed. The average of these resulting coexpression values is called the *coordinated expression of a complex* throughout this study. Computing the coordinated expression for each protein complex in the InWeb network for each (normal) tissue results in a tissue - coexpression association matrix providing information about the strength of coordinated expression of a hub protein and its interactors in any protein complex of the network in any normal tissue.

**Detecting coordinated expression in disease related protein complexes** Lage and colleagues [43] carried out a study in which they could detect a significant tendency for disease genes and complexes to be over-expressed in normal tissues where defects cause pathology. In this study we want to analyze to what extent disease complexes are well coordinated in expression in normal disease-relevant tissues. We hypothesize that a protein complex, in which the hub protein is confirmed to be involved in a certain disease, would be well coordinated in expression in normal tissue where mutations cause disease phenotype.

## 8.6 A pipeline of gene prioritization tools for identifying novel disease genes

In our review presented in *Study 1*, we wanted to clarify gene prioritization to biologists or clinicians through a guide of 19 gene prioritization tools. As a follow up, we have initiated *Study 2*, critically assessing the gene prioritization tools' performance from the previous study through a benchmark of recently discovered disease-gene associations. Taking one step further, we have recently started a new study to assess some of the gene prioritization tools from our previous studies in a pipeline to discover novel disease genes. The goal of this study is to assess the usefulness of a systematic combination of different gene prioritization tools for suggesting novel disease genes leading from a genome-wide ranking to determine a set of promising candidate genes towards a ranking of promising candidate genes resulting in a short list of potential novel disease genes. We are convinced that this study could be a benefit for biologists and clinicians in their research.

We have chosen the best performing genome-wide gene prioritization tools (CANDID [39] and PINTA (*Study 5*)) and the best performing non-genome-wide gene prioritization tools (Endeavour [50] and Genedistiller [15]) from *Study 2* to obtain optimal results. For this study, we have modified the implementation of PINTA by replacing the expression data with training genes as input: instead of

using expression values for each gene, only training genes get a 1, while all other genes get a 0.

First, we define a two-layer structure: on the first layer, tools are located performing genome-wide gene prioritization (CANDID, PINTA). We run these tools on a disease under study to obtain a genome-wide ranking list of genes. To minimize the number of false positives, we merge these two ranking lists by using order statistics and take the top 10% (approx. 2000 genes) from this list as input for the second layer. On the second layer, tools are located performing non-genome-wide gene prioritization (Endeavour, Genedistiller). The genes from the first merged ranking list resemble the test genes for the tools on the second layer which will be prioritized again. The resulting two ranking lists are merged using order statistics to reduce again the number of false positives among the high ranked genes. Taking the top 10% of the second merged ranking list should represent a stable and reliable ranking list of genes that could be involved in the disease under study.

Before we can start doing these experiments, we have to determine diseases for which we want to predict novel disease genes using our two-layer concept of gene prioritization tools. For the diseases we have to collect keywords and known disease genes (as training genes) by experts, because PINTA and Endeavour need training genes, whereas CANDID and Genedistiller need keywords as input data. To retrieve high quality keywords and known disease genes, we have started to contact collaborators in the UZ Leuven (Bernard Thienpont and Jeroen Breckpot for the study on Congenital Heart Defect (CHD), and Paul Brady for the study on Congenital Diaphragmatic Hernia (CDH)). We are now aiming at getting more data for other diseases from other clinicians and researchers to set up a solid set of diseases.

The aim of this study is to obtain a short list of potential novel disease genes by the systematic combination of different gene prioritization tools based on a two-layer concept. The suggested novel disease genes resulting from this study could then be investigated further in a follow-up study for a causal effect in a certain disease.

## 8.7 Conclusion

We have introduced an approach for gene prioritization by replacing prior knowledge about the biological process with experimental data on differential gene expression between affected and healthy individuals. Considering the problem from the perspective of a gene or protein network, we assess a candidate gene by considering the level of differential expression in its neighborhood under the assumption that strong candidates will tend to be surrounded by differentially expressed neighbors.

In our first study the algorithm was based on a notion of soft neighborhood where

each gene is given a contributing weight, which decreases with the distance from the candidate gene on the protein network. To account for multiple paths between genes, we have defined the distance using the Laplacian exponential diffusion kernel. We have scored candidates by aggregating the differential expression of neighbors weighted as a function of distance.

The follow up study has advanced this work extensively by applying machine learning approaches based on random walk models. We have explored three different random walk based strategies that were benchmarked on 40 publically available knockout experiments in mice. Results showed that the best results were obtained using the Heat Kernel Diffusion Ranking leading to an average ranking position of 8 out of 100 genes, an AUC value of 92.3%.

After that we have developed PINTA, a freely accessible web resource implementing the concepts from the previous studies. PINTA is available for some prominent model organisms beside human and various array platforms are supported and can identify promising candidate genes using network-based machine learning approaches even if no knowledge is available about the disease or phenotype.

# Bibliography


[1] Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti, and et al. Mint, the molecular interaction database: 2009 update. *Nucleic Acid Research (Database Issue)*, 38:D532–539, 2010.

[2] Zhang A. *Protein Interaction Networks: Computational Analysis*. Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA, 2009.

[3] Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, and et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[4] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, and et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 101(16):6062–7, 2008.

[5] Barabasi AL and Oltvai ZN. Network biology: understanding the cell's functional organization. *Nature Reviews: Genetics*, 5:101–113, 2004.

[6] Francisco AP, Goncalves JP, Madeira SC, and Oliveira AL. Using personalized ranking to unravel relevant regulations in the saccharomyces cerevisiae regulatory network. In *Proceedings of Jornadas de Bioinformatica. Lisbon*, pages 3–6, 2009.

[7] Aranda B, Achuthan P, Alam-Faruque Y, ArmeanI, Bridge A, and et al. The intact molecular interaction database in 2010. *Nucleic Acids Research*, doi:10.1093/nar/gkp878, 2009.

[8] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, and et al. Biogrid: A general repository for interaction datasets. *Nucleic Acid Research (Database Issue)*, 34:D535–539, 2006.

[9] Wolfe C, Kohane IS, and Butte AJ. Systematic survey reveals general applicability of guilt-by-association within gene coexpression networks. *BMC Bioinformatics*, 6:227, 2005.

[10] The Personalized Medicine Coalition. Personalized medicine: An introduction. *http://www.personalizedmedicinecoalition.org/*.

[11] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[12] Cheng D, Knox C, Young N, Stothard P, Damaraju S, and et al. Polysearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research (Web server issue)*, 36:W399–405, 2008.

[13] Forler D, Köcher T, Rode M, Gentzel M, Izaurralde E, and et al. An efficient protein complex purification method for functional proteomics in higher eukaryotes. *Nature Biotechnology*, 21(1):89–92, 2003.

[14] Pinkel D, Segraves R, Sudar D, Clark S, Poole I, and et al. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2):207–211, 1998.

[15] Seelow D, Schwarz JM, and Schuelke M. Genedistiller - distilling candidate genes from linkage intervals. *PLoS ONE*, 3:e:3874, 2008.

[16] de Bie T, Tranchevent L, van Oeffelen L, and Moreau Y. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–132, 2007.

[17] Watts DJ and Strogatz SH. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[18] Klipp E, Herwig R, Kowald A, Wierling C, and Lehrach H. *Systems Biology in Practice*. Wiley-VCH, Weinheim, 2005.

[19] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, and Barabasi AL. Hierachical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[20] Adie EA, Adams RR, Evans KL, Porteous DJ, and Pickard BS. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22:773–774, 2006.

[21] Phizicky EM and Fields S. Protein-protein interactions: methods for detection and analysis. *Microbiological Reviews*, 59:94–123, 1995.

[22] Fouss F, Pirotte A, Renders JM, and Saerens M. Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19:355–369, 2007.

[23] Fouss F, Francoisse K, Yen L, Pirotte A, and Saerens M. An experimental investigation of graph kernels on collaborative recommendation and semi-supervised classification. *Submitted for publication*, 2008.

[24] Fouss F, Yen L, Pirotte A, and Saerens M. An experimental investigation of graph kernels on a collaborative recommendation task. In *IEEE International Conference on Data Mining (ICDM)*, pages 863–868, 2006.

[25] Soldner F, Hockemeyer D, Beard C, Gao Q, Bell GW, and et al. Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell*, 136(5):964–977, 2010.

[26] Miller G. Human lymphoblastoid cell lines and epstein-barr virus: a review of their interrelationships and their relevance to the etiology of leukoproliferative states in man. *Yale Journal of Biology and Medicine*, 43(6):358–384, 1971.

[27] Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, and et al. Bind - the biomolecular interaction network database. *Nucleic Acid Research*, 29(1):242–245, 2001.

[28] Ge H. Upa, a universal protein array system for quantitative detection of protein-protein, protein-dna, protein-rna and protein-ligand interactions. *Nucleic Acids Research*, 28(2):d3, 2000.

[29] Zhu H and Snyder M. Protein chip technology. *Current Opinion in Chemical Biology.*, 7(1):55–63, 2003.

[30] Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, and et al. Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, 8(3):R39, 2007.

[31] Park IH, Zhao R, West JA, Yabuuchi A, Huo H, and et al. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature*, 451(7175):141–146, 2008.

[32] Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, and et al. Ecocyc: a comprehensive database resource for escherichia coli. *Nucleic Acid Research(Database issue)*, 33:D334–337, 2005.

[33] Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, and et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27(2):199–204, 2009.

[34] Chen J, Xu H, Aronow BJ, and Jegga AG. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, 8:392, 2007.

[35] Jozefczuk J, Prigione A Chavez L, and Adjaye J. Comparative analysis of human embryonic stem cell and induced pluripotent stem cell-derived hepatocyte-like cells reveals current drawbacks and possible strategies for improved differentiation. *Stem cells and development*, doi:10.1089/scd.2010.0361, 2011.

[36] Quackenbush J. Genomics. microarrays - guilt by association. *Science*, 302(5643):240–241, 2003.

[37] Shawe-Taylor J and Cristianini N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[38] Yu J, Vodyanik MA, Smuga-Otto K, Antosievicz-Bourget J, Frane JL, and et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–1920, 2007.

[39] Hutz JE, Kraja AT, McLeod HL, and Province MA. Candid: a flexible method for prioritizing candidate genes for complex human traits. *Genetic Epidemiology*, 32:779–790, 2008.

[40] Dimos JT, Rodolfa KT, Niakan KK, Weisenthal LM, Mutsumoto H, and et al. Induced pluripotent stem cells generated from patients with als can be differentiated into motor neurons. *Science*, 321(5893):1218–1221, 2008.

[41] Zhao JT, Cowley MJ, Lee P, Birzniece V, Kaplan W, and et al. Growth hormone regulation of lipid metabolism in adipose tissue: a microarray analysis in hypopituitary men. 2001.

[42] Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, and et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316, 2007.

[43] Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, and et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *PNAS*, 105(52):20870–5, 2008.

[44] Raman K and Chandra N. Systems biology. *Resonance*, 15:131–153, 2010.

[45] Takahashi K, Saitoh S, and Yanagida M. Application of the chromatin immunoprecipitation method to identify in vivo protein-dna associations in fission yeast. *Science Signaling*, 2000:pl1, 2000.

[46] Brown KR and Jurisica I. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, 2005.

[47] Lovasz L. Random walks on graphs: A survey. In *Combinatorics, Paul Erdös is eighty, Vol. 2 (Keszthely, 1993), volume 2 of Bolyai Soc. Math. Stud*, pages 353–397, 1993.

[48] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, and et al. The database of interacting proteins: 2004 update. *Nucleic Acid Research (Database Issue)*, 32:D229–251, 2004.

[49] Sie L, Loong S, and Tan EK. Utility of lymphoblastoid cell lines. *Journal of Neuroscience Research*, 87:1953–1959, 2009.

[50] Tranchevent L, Barriot R, Yu S, Van Vooren S, Van Loo P, and et al. Endeavour update: a web resource for gene prioritization in multiple species. *Nucleic Acids Research, Web Server issue*, 36(1):377–384, 2008.

[51] Jorde LB. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 10(10):1435–1444, 2000.

[52] Hartwell LH, Hopfield JJ, Leibler S, and Murray AW. From molecular to modular cell biology. *Nature*, 402:c47–c52, 1999.

[53] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, and et al. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acid Research (Database Issue)*, pages D412–416, 2009.

[54] Bakay M, Wang Z, Melcon G, Schiltz L, Xuan J, and et al. Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of rb-myod pathways in muscle regeneration. *Brain*, 129(Pt 4):996–1013, 2006.

[55] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, and et al. Kegg for linking genomes to life and the environment. *Nucleic Acid Research(Database issue)*, 36:D480–D484, 2008.

[56] Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, and et al. A global map of human gene expression. *Nature Biotechnology*, 28(4):322–324, 2010.

[57] Oti M and Brunner HG. The modular nature of genetic diseases. *Clinical Genetics*, 71:1–11, 2007.

[58] Saerens M, Fouss F, Yen L, and Dupont P. The principal components analysis of a graph, and its relationships to spectral clustering. In *Proceedings of the 15th European Conference on Machine Learning (ECML 2004). Lecture Notes in Artificial Intelligence, Vol. 3201, Springer-Verlag, Berlin*, pages 371–383, 2004.

[59] Schena M, Shalon D, Davis RW, and Brown PO. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[60] Urbanek M, Legro RS, Driscoll DA, Azziz R, and Ehrmann DA. Thirty-seven candidate genes for polycystic ovary syndrome: strongest evidence for linkage is with follistatin. *Proceedings of the National Academy of Sciences of the United States of America*, 96(15):8573–8578, 1999.

[61] Urbanek M, Sam S, Legro RS, and Dunaif A. Identification of a polycystic ovary syndrome susceptibility variant in fibrillin-3 and association with a metabolic phenotype. *Journal of Clinical Endocrinology and Metabolism*, 92(11):4191–4198, 2007.

[62] Yanagida M. Functional proteomics: current achievements. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Science*, 771:89–106, 2002.

[63] Zhu M and Zhao S. Candidate gene identification approach: Progress and challenges. *International Journal of Biological Sciences*, 3(7):420–427, 2007.

[64] Spillman MA, Manning NG, Dye WW, Sartorius CA, Post MD, and et al. Tissue-specific pathways for estrogen regulation of ovarian cancer growth and metastasis. *Cancer Research*, 70(21):8927–36, 2010.

[65] Ramocki MB, Bartnik M, Szafranski P, Kolodziejska KE, Xia Z, and et al. Recurrent distal 7q11.23 deletion including hip1 and ywhag identified in patients with intellectual disabilities, epilepsy, and neurobehavioral problems. *American Journal of Human Genetics*, 87(6):857–865, 2010.

[66] Marazita ML, Murray JC, Lidral AC, Arcos-Burgos M, Cooper ME, and et al. Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35. *American Journal of Human Genetics*, 75(2):161–173, 2004.

[67] Kandola N, Cristianini N, and Shawe-Taylor J. Learning semantic similarity. *Advances in Neural Information Processing Systems*, pages 657–664, 2002.

[68] Miyake N, Kosho T, Mizumoto S, Furuichi T, Hatamochi A, and et al. Loss-of-function mutations of chst14 in a new type of ehlers-danlos syndrome. *Human Mutation*, 31(8):966–974, 2010.

[69] Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, and et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Research*, 34(10):3067–3081, 2006.

[70] Turton NJ, Judah DJ, Riley J, Davies R, Lipson D, and et al. Gene expression and amplification in breast carcinoma cells with intrinsic and acquired doxorubicin resistance. *Oncogene*, 20(11):1300–1306, 2001.

[71] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, and Botstein D. A bayesian framework for combining heterogeneous data sources for gene function prediction. *Proceedings of the National Academy of the Science of the USA*, 100(14):8348–8353, 2003.

[72] Baldi P and Long AD. A bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.

[73] Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, and et al. The mips mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

[74] Puigserver P. Tissue-specific regulation of metabolic pathways through the transcriptional coactivator pgc1-alpha. *International Journal of Obesity*, 29(Suppl1):S5–9, 2005.

[75] Albert R and Barabasi AL. Statistical mechanics of complex networks. *Reviews of Modern Physiks*, 74:47–97, 2002.

[76] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, and et al. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

[77] Sharan R, Ulitsky I, and Shamir R. Network-based prediction of protein function. *Molecular Systems Biology*, 3(88), 2007.

[78] Kondor RI and Lafferty JD. Diffusion kernels on graphs and other discrete structures. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publisher Inc*, pages 315–322, 2002.

[79] Sapolsky RJ, Hsie L, Berno A, Ghandour G, Mittmann, and et al. High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays. *Genetic Analysis: Biomolecular Engineering*, 14(5-6):187–192, 1999.

[80] Schilsky RL. Personalized medicine in oncology: the future is now. *Nature Reviews. Drug Discovery*, 9(5):363–366, 2010.

[81] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, and et al. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24:537–544, 2006.

[82] Ahn S, Wang RT, Park CC, Lin A, Leahy RM, and et al. Directed mammalian gene regulatory networks using expression and comparative genomic hybridization microarray data from radiation hybrids. *PloS Computational Biology*, 5(6):e1000407, 2009.

[83] Dudoit S, Yang YH, Callow MJ, and Speed TP. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, 12:111–139, 2002.

[84] Jones S and Thornton JM. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93:13–20, 1996.

[85] Köhler S, Bauer S, Horn D, and Robinson PN. Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82(4):949–958, 2008.

[86] Zhu S, Chen PC, and Yen J. Multi-layer bayesian network for variable-bound inference. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, volume 2, pages 553 –559, 2008.

[87] Ito T, Shimbo M, Kudo T, and Matsumoto Y. Application of kernels to link analysis. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 586–592, 2005.

[88] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, and et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[89] Lee T, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, and et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[90] Pers TH, Hansen NT, Lage K, Koefoed P, Dworzynski P, and et al. Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. *Genetic Epidemiology*, in press, 2011.

[91] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, and et al. Human protein reference database–2009 update. *Nucleic Acid Research (Database Issue)*, 37:D767–772, 2009.

[92] Alon U. Biological networks: the tinkerer as an engineer. *Science*, 301:1866–1867, 2003.

[93] Spirin V and Mirny LA. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):2123–2128, 2003.

[94] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, and et al. String: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–261, 2003.

[95] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, and et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.

[96] Sadee W and Dai Z. Pharmacogenetics/genomics and personalized medicine. *Human Molecular Genetics*, 14(2):R207–214, 2005.

[97] Cui X and Churchill GA. Statistical tests for differential expression in cdna microarray experiments. *Genome Biology*, 4:210, 2003.

[98] Nishimura Y, Martin CL, Vazquez-Lopez A, Spence SJ, Alvarez-Retuerto AI, and et al. Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Human Molecular Genetics*, 16(14):1682–1698, 2007.

[99] Yoshida Y and Yamanaka S. Recent stem cell advances: Induced pluripotent stem cells for disease modeling and stem cellŰbased regeneration. *Circulation*, 122:80–87, 2010.

[100] Yoshida Y, Makita Y, Heida N, Asano S, Matsushima A, and et al. Posmed (positional medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Research (Web server issue)*, 37:W147–152, 2009.

[101] Bar-Joseph Z. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–2503, 2004.

[102] Yao Z, Jaeger JC, Ruzzo WL, Morale CZ, Emond M, and et al. A marfan syndrome gene expression phenotype in cultured skin fibroblasts. *BMC Genomics*, 8:319, 2007.

# List of publications

## Journal papers

- **Nitsch D**, Tranchevent L, Goncalves JP, Vogt JK, Madeira SC, Moreau Y (2011). PINTA - A web server for network-based gene prioritization from expression data. *Nucleic Acids Research* vol. 39 (suppl 2): W334-W338.
- **Nitsch D**, Goncalves JP, Ojeda F, de Moor B, Moreau Y (2010). Candidate Gene Prioritization by Network Analysis of Differential Expression using Machine Learning Approaches. *BMC Bioinformatics* 11:460.
- Tranchevent L*, Capdevila FB*, **Nitsch D***, De Moor B, De Causmaecker P, Moreau Y (2011). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics* 12(1):22-32. Published online March 21, 2010 doi:10.1093/bib/bbq007. *Equally contributed.
- **Nitsch D**, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y (2009) Network Analysis of Differential Expression for the Identification of Disease-Causing Genes. *PLoS ONE 4*(5): e5526.

### Manuscripts submitted or in preparation

- Fassbender A, Verbeeck N, **Nitsch D**, Kyama CM, Bokor A, Vodolazkaia A, Peeraer K, Tomassetti C, Meuleman C, Gewaert O, van de Plas R, de Moor B, Moreau Y, Waelkens E, D'Hooghe T (2011). Combined Microarray and Proteomic analysis of eutopic endometrium of women with and without endometriosis. Internal Report, 11-77, ESAT-SISTA, K.U.Leuven (Leuven, Belgium). Submitted.
- **Nitsch D***, Tranchevent L*, Capdevila FB*, Devriendt K, De Moor B, De Causmaecker P, Moreau Y (2011). A Critical Assessment of Gene Prioritization. Internal Report, 11-15, ESAT-SISTA, K.U.Leuven (Leuven, Belgium). *Equally contributed

- Capdevila FB*, **Nitsch D***, Tranchevent L*, De Moor B, Moreau Y, De Causmaecker P (2011). A Pipeline through Gene Prioritization methods for identifying novel disease genes. Internal Report, 11-16, ESAT-SISTA, K.U.Leuven (Leuven, Belgium). *Equally contributed

- **Nitsch D**, Pers TH, Brunak S, Moreau Y (2011). Identifying tissue-specificity of protein-complexes based on a global map of human expression data. Internal Report, 11-96, ESAT-SISTA, K.U.Leuven (Leuven, Belgium).

- Pers TH, **Nitsch D**, Moreau Y, Brunak S, Sørensen TIA (2011). Protein complex analysis associates SNPs in ASIP with extreme overweight. Internal Report, ESAT-SISTA, K.U.Leuven (Leuven, Belgium).

## International conferences: oral presentations

- **Nitsch D**, Tranchevent LC, Moreau Y (2011). Network-based gene prioritization from expression data by diffusing through protein interaction networks. *ISMB/ECCB 2011*, July 17-19 2011, Vienna, Austria.

- **Nitsch D**, Tranchevent LC, Moreau Y (2011). Machine learning approaches for network-based gene prioritization from expression data. *5th International Workshop on Machine Learning in Systems Biology*, July 20-21 2011, Vienna, Austria.

- **Nitsch D**, Tranchevent LC, Moreau Y (2011). PINTA - a web server for network-based gene prioritization from expression data. *Network Biology SIG, ISMB 2011*, July 15 2011, Vienna, Austria.

- **Nitsch D**, Goncalves JP, Ojeda F, de Moor B, Moreau Y (2010). Prioritizing Candidate Genes by Network Analysis of Differential Expression using Machine Learning Approaches. *RECOMB 2010*, August 12-15 2010, Lisbon, Portugal.

- **Nitsch D**, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y (2009) Disease gene identification by network analysis of differential Expression. *23rd EURO Conference on Operational Research, Stream: Computational Biology, Bioinformatics and Medicine*, July 5-8 2009, Bonn, Germany.

- **Nitsch D**, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y (2009) Disease gene identification by network analysis of differential expression - an update. *5th Bioptrain Workshop*, June 10-12 2009, Florence, Italy.

- **Nitsch D**, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y (2009) Disease gene identification by network analysis of differential expression. *Genomic Disorders Conference on Genomic Variation in Health and Disease*, March 9-11 2009, Wellcome Trust Center, Hinxton, UK.

- **Nitsch D**, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y (2009) Disease gene identification by network analysis of differential expression. *4th Bioptrain Workshop*, January 11-3 2009, Innsbruck, Austria.

- **Nitsch D**, Moreau Y (2008) Ab initio gene prioritization. *Mini EURO Conference on Computational Biology, Bioinformatics and Medicine*, September 15-17 2008, Rome, Italy.

# Curriculum vitae

Daniela Börnigen, nee Nitsch, was born in Lübeck (Germany) on March 7, 1981. She started studying Media Computer Science at the University of Applied Sciences Wedel (Germany) in 2000 and obtained her diploma in 2005. In 2004 she went abroad to continue her studies at City Varsity in Cape Town (South Africa). Her thesis on medical image processing was carried out in collaboration with Philips Medical Systems in Hamburg (Germany). Between 2005 and 2007 she completed her master of computer science at the University of Ulm (Germany). She carried out her thesis project on Boolean Networks at the Institute of Neural Information Processing under the supervision of Dr. Hans Kestler and Prof. Dr. Günther Palm. In September 2007, she started her PhD at the Katholieke Universiteit Leuven under the supervision of Prof. Dr. Yves Moreau, Prof. Dr. Louis Wehenkel, Prof. Dr. Koenraad Devriendt and Prof. Dr. Bart De Moor.