HETEROGENEOUS INFORMATION SOURCES FOR BIOINFORMATICS: INTEGRATION METHODOLOGY, SEARCH ALGORITHMS AND CASE STUDIES

Francisco BONACHELA CAPDEVILA

Supervisor:
Prof. dr. P. De Causmaecker

Members of the Examination Committee:
Prof. dr. P. Igodt, chair
Prof. dr. H. Deckmyn
Prof. dr. H. Pottel
Prof. dr. J. Garibaldi, (University of Nottingham, UK)
Prof. dr. Y. Moreau
Prof. dr. D. Pelta (Universidad de Granada, Spain)
Prof. dr. Ellen Decaestecker

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Science

October 2012

© 2009 Katholieke Universiteit Leuven, Groep Wetenschap & Technologie, Arenberg Doctoraatsschool, W. de Croylaan 6, 3001 Heverlee, België

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

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

ISBN number 978-90-8649-562-7
Legal depot number D/2012/10.705/79

PREFACE

This thesis summarizes the work that I have performed in the CODeS and ITEC groups in KU Leuven campus Kortrijk thanks to the support of the BIOPTRAIN program and the KU Leuven *campus impuls financiering*. During these last years, I have had the opportunity of working in a very stimulating atmosphere and interacting with researchers from many different countries who brought different views and perspectives to my own work. I could also learn from worldwide known researchers and visit other laboratories and colleagues all around Europe. After all these years, if I look back, I can say that taking the step of coming to Kortrijk was one of the most right decisions I have ever made.

I would like to deeply thank my promotor Prof. Patrick De Causmaecker, first for his courage to apply his expertise into a new field of research, second for giving me the opportunity of joining his team and third for all the unconditional support that I have felt all these years both in the scientific and the personal sides.

I am also grateful to Prof. Hans Deckmyn. He helped me from the very beginning to know more deeply the biological side of my research and he introduced me to very talented researchers in an international consortium where I could apply our cluster approach. I would also like to thank Dr. Katleen Broos for her comments, her work and her feedback on the results of our work. Without them, this work would be less complete.

My gratitude also goes to Prof. Ellen De Caestecker for her support during the last months of my Ph.D., for her interdisciplinary enthusiasm and for introducing us to the world of *daphnia*, where there is still so many interesting things to do.

I also thank Prof. Yves Moreau for his help and reviews in gene prioritization that made my approach more solid. Prof. Hans Pottel gave me hints about how to design experiments. Prof. Catherine Laprise was invaluable analyzing genes and Prof. Djamal Rebaine offered me new insights and the possibility of expanding them.

I am also grateful to Prof. Jonathan Garibaldi, responsible of the program BIOPTRAIN that brought me to Kortrijk and all the members of this group for so many interesting research meetings. They are Daniel, Enrico, Pawel, Andrea, Matthieu, Marc, Linda, Aleksandra, Daniela and Pooja. I would like to have few particular words to Léon-Charles Tranchevent, whose work has been always a reference to me and whose words and comments have considerably helped me during all my research.

I also thank Prof. David Pelta and Prof. José Luis *Curro* Verdegay. I started with them a journey that after seven years, it is still moving forward.

Most of the important moments of my Ph.D. have passed in Kortrijk, continuously in touch with people of ITEC and CODeS, with Geraldine, Piet and Wim as heads and the rest of colleagues: Hans, Stefan, Carmen, Karolien, Caroline, Wilfried, Joke, Trevor, Brendan

Antoine, Frederik, Igor, Ruben, Maribel, Mieke and Kelly. It has been a pleasure to work with all of them.

I would like to dedicate few words to some colleagues of CODeS with whom I have shared years of research and friendship. Stefaan and Tommy have made the time at the office fly and Kuchi, Ahmet, Bidzina and Lei have helped me different perspectives of life.

KULAK is really much more than the research team where I have been working. I would like to mention the incredible Chris De Paepe, a true conquistador, always with a smile in his mouth and encyclopedic knowledge of Spanish and the Hispanic world. Virginie Coucke helped me so much in my first months that without her, I would probably have not survived so long in Kortrijk. Paul Igodt welcomed me my first day and made me feel comfortable far away from home. Thanks also to Brecht & Brecht, Piet and Jan. To Griet and her permanent smile and to so many colleagues that speeded the train to Gent up every day: Stijn, Bart, Eva and Tim.

Joepie and Quinten, Quinten and Joepie. They always helped this Spanish visitor to understand (as much as possible) the country that welcomed him. One of them discovered the painful honey. The other one never stopped cranking it up a notch. They know who is who.

A very big thanks to my friends back in Spain (Joaquín, Javier, Ramón, Antonio, David, Fran...) *¡Gracias chavales por estar ahí!*. To my family in Catalonia and Andalusia with special words for Ramón, Josep and Roger *¡Estius inoblidables!* To my brother and to the countless hours that we have spent together, laughing, talking, playing and living. And of course to my parents who, with effort and without reserve, taught me to be the person that I am today. *¡Gracias papa!¡Gràcies mama!*

And finally, I could not finish this preface without naming Ana, her smile and her infinite patience. *Tienes el cielo ganado*. Thank you.

Francisco Bonachela Capdevila

Kortrijk, June 2012

ABSTRACT

Identifying the genetic basis associated with Mendelian disorders or complex phenotypes is essential in human genetics in order to design more effective and eventually to better understand the molecular mechanisms behind these genetic disorders.

Usually, a list of candidates is obtained in a high-thoughput experiment, such as a genomewide association study. This set of genes (either a chromosomal region or a list of genes scattered in the genome) is usually not small enough to easily undertake a manually one-by-one validation and therefore a selection of the putative most interesting genes is needed. This problem has been named *gene prioritization* and in the last years, several computing based approaches have been proposed to cope with it. This thesis presents a work on gene prioritization.

The first part of this text thoroughly reviews the web based gene prioritization tools that can be freely used by any user. We describe seventeen tools and we stress their similarities and differences with the aim to help the user to choose the most appropriate one for his type of data. We have also reviewed the bibliography associated with these tools in search of validations and tool performance comparisons and we have finally set up a website where this information and regular updates are stored. In the last two years, the number of tools described in the website has almost doubled.

Furthermore, we have developed a performance review among gene prioritization tools, both using the whole genome as starting candidate set or a limited one. We have compared individual results with the combination of the tools and finally we have completed our review with the combination of the best performance gene prioritization tools in our benchmark in three real life experiments. All the expertise gathered in our complete review has been used to find new candidate genes involved in congenital heart disease, congenital diaphragmatic hernia and asthma.

Finally, we propose the use of cluster analysis as a preprocessing step of gene prioritization approaches that use training genes to lead the prioritization. We claim that the automatic selection of a homogenous training set produces more accurate rankings than the expert selected ones. To this purpose, we have applied a transactional clustering algorithm, CLOPE, to two different gene prioritization tools: Endeavour and Genedistiller.

CONTENTS

Chapter 1

Introduction

1.1 HUMAN GENETICS

The morphological and functional unity of every living being is called *cell*. These (usually) tiny elements (the word cell originates from the latin *cellam*, small room) are the smallest elements generally considered alive. All vital functions in any living being depend on how the cells work and interact.

Eukaryotic cells, the building blocks of complex organisms, including human beings, are contained within a membrane that also forms a number of organelles, each one with a particular function, floating in a gel-like substance called cytoplasm. Besides these organelles, there is the core of the cell, the cell nucleus. There, surrounded by a second membrane, is the *sancta sanctorum* of the cell and by extension, of life: the DNA. Figure 1 represents in a simplified manner the location of the DNA within a cell and its physical structure.

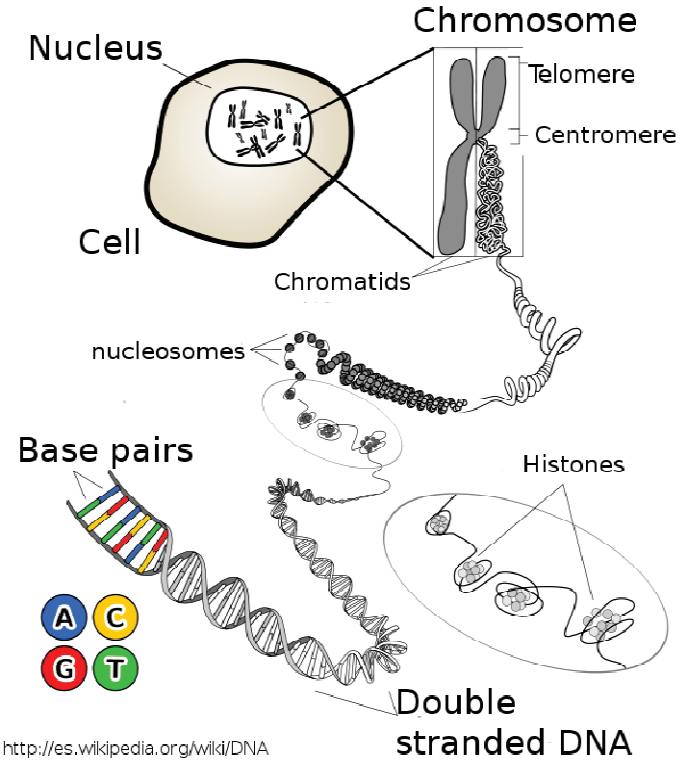


Figure 1.- Simplified representation of the location of the DNA within a cell and its physical structure

The reason why we are like we are can be found in a long and thin macromolecule called DNA (desoxyribonucleic acid). DNA is a polymer of nucleotides and each nucleotide consists of a sugar ring (the same one for all DNA nucleotides), a phosphate group (the same one for all DNA nucleotides) and a nitrogenous base, which can be either adenine (A), guanine (G), thymine (T) or cytosine (C). Therefore, the relative position of every of these nitrogenous bases in the long polymer of the DNA will make the difference between two DNA strings.

The human genome contains about 3 billion nucleotides in 23 chromosomes. The DNA contains information needed to build the proteins supporting the human body (structural proteins), protecting it from external hazards (antibodies) and controlling the metabolism (enzymes) and any other functions (signal transduction). This information, in the form of genes, follows the so-called *central dogma of the molecular biology*. It is first transcribed from DNA to RNA (a different type of nucleic acid), which leaves the nucleus and, with the help of the organelles called ribosomes, it is translated into proteins (which are themselves polymers of amino acids).

It has been estimated that there are between 20000 and 25000 protein coding genes in the human genome. However, the number of proteins that can be built is much higher, due to processes like alternative splicing, where pieces of RNA produced by transcription of a gene can be connected in different manners and therefore different products can be obtained.

Therefore, the original sequence of As, Cs, Gs and Ts of the DNA strands controls how our body is built and how it works, why we are different and, for example, our susceptibility to various diseases.

DNA comes in a double string of nucleotides. The bases adenine and thymine chemically attract each other by the formation of hydrogen bridges, as do cytosine and guanine. At the corresponding position, when in one strand of DNA we find a T, in the other string there will be an A (and vice versa) and if we find a G, in the other string there will be a C and vice versa: the two strings are complementary.

Several types of genetic alterations have been found, which can account for both human traits variability and genetic diseases:

- Single Nucleotide Polymorphism.- “SNPs are single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater” [1]. A single base alteration with a frequency less than 1% is called mutation. In both cases, the alteration can modify the probability of suffering from a particular disease.
- Copy-number variations.- CNVs are chromosomal structural variations resulting in an abnormal number of copies of one or more regions of the DNA due to both deletion or duplication. In the last years, the number of complex diseases linked to an abnormal number of copies of chromosomic regions is increasing [2].

- Structural rearrangement or translocation.- A complete region is moved from one chromosome to another. In this case, there is no gain or loss in the total amount of genetic information content but the order of the nucleotides changes. Subtypes of acute myelogenous leukemia and chronic myelogenous leukemia are caused by chromosomal translocations [3, 4].
- Epigenetic modifications.- Involves any changes in gene expression not caused by changes in the underlying DNA sequence.

Understanding the genetic basis of the diseases is an essential aim of human genetics in general and medical genetics in particular and it will be the subject this work will focus on. Knowing the disease mechanisms leads to more effective therapies and treatments.

Humans have known for thousands of years that health is affected by heredity [5]. However, we usually name the works of the Czech monk Gregor Mendel in the second half of 19th century as the starting point of genetics. He studied inheritance in plants and in 1865 he presented his paper *Versuche über Pflanzenhybriden* ("Experiments on Plant Hybrids"), where the inheritance patterns of certain traits in pea plants were described. However, it was not until the beginning of the 20th century that his work was acknowledged by the scientific community [6, 7] and the first studies relating diseases with heredity were launched [8].

During the 20th century, a series of discoveries made genetics knowledge boost. In 1910 Thomas Hunt Morgan placed genes on chromosomes. In 1941 George Beadle and Edward Tatum identified that proteins originate from genes, making a first step towards the definition of the central dogma of the molecular biology. In 1944 the Avery-MacLeod-McCarty experiment identified DNA as the genetic material. In 1950 Erwin Chargaff proved that the four nucleotides are present in numbers based on a general rule. In 1953 James Watson and Francis Crick, based on previous work of Rosalind Franklin proposed the double helix model for the DNA chain.

With the improvement of sequencing techniques, starting with Sanger's sequencing work in 1977, many genomes of different organisms have been sequenced, contributing to the development of genetics but the cornerstone of human genetics happened in the beginning of the 21st century, when the human genome was sequenced [9].

The sequencing of the human genome opened the door to the high-throughput experiments where thousands of experiments can be performed in parallel. A well known example is microarrays, where a chip containing a collection of samples of DNA is used to measure the expression of thousands of genes simultaneously.

1.2 BIOINFORMATICS

In parallel to the development of genetics, the second half of 20th century witnessed a boost in computing. The ever-increasing computer power of integrated circuits based computers started to be used to mine the huge databases that high-throughput

experiments were creating. For instance GenBank [10], an open access sequence database which contains nucleotide sequences and their protein translations has doubled its records every 18 months since 1982. PDB [11], the protein data bank, a repository for the 3-D structural data of large biological molecules such as proteins and nucleic acids, has passed from 13597 structures in 2000 to 82809 in 2012.

In the last years, plenty of computational solutions have been implemented to satisfy the requirements of geneticists: homology searching (e.g. BLAST), sequence alignment (e.g. ClustalW), phylogenetics (e.g. PHYLIP), functional patterns (e.g. HMMER), gene prediction (e.g. GenScan), regulatory region analysis (e.g. MatInspector), RNA structure (e.g. UniFold), protein structure (e.g. JPred)…

The number of databases storing biologically related information has also enormously increased in the last decades. First data repositories were set up at the end of the 60s [12, 13]. Nowadays, the number of publicly available databases with biological information exceeds thousand [14].

1.3 GENE PRIORITIZATION

The concept of gene prioritization can be defined as the sorting of a set of genes based on their characteristics in order of relevance with respect to a particular biological process. When this process is applied to a gene list containing a gene of interest, the user expects this gene to be ranked in the first positions. Gene prioritization can be seen as a natural continuation to classic biological approaches such as linkage analysis or genome-wide association studies which usually return an output containing hundreds of genes when are applied to hunt a particular gene of interest (involved in a disease, for instance) [15, 16].

This concept was first introduced ten years ago by Pérez-Iratxeta, who already described an approach to this problem [17]. This approach takes advantage of both the progress made in computational development and the large amount of genomic data publicly available. Since then, in these ten last years, many approaches have been developed, among which some have been implemented into web applications and eventually validated [18–46]. These tools use different strategies, different inputs and outputs and the databases where the gene information is collected are also diverse. However, these tools agree on the *guilt-by-association* concept: the most promising candidates to pursue the research with a e.g. knock-down experiment will be the ones that are similar to the genes that are already known to be linked to the biological process of interest [47–49].

One of the possible strategies used during the gene prioritization process involves the analysis of two sets of genes. The *candidate* set contains the list of genes that will be eventually ranked and the *training* set of genes includes a list of genes, usually associated or linked to the disease, from which a profile will be retrieved in order to rank the candidate genes. The final ranking will depend on the similarity of the candidate genes with the training genes profile and the measure of this similarity will depend on a set of databases.

The selection of the three factors (candidate set, training set and databases) are critical steps upon which the final accuracy of the ranking fully depends, and in some tools, these critical decisions must be taken by the final user.

1.3.1 CANDIDATE SET

Any gene set can be ranked. However, as long as the set contains a gene of interest, it is natural to admit that the smaller the set is, the easier the gene of interest will be ranked in high positions. The full genome can be used as candidate set (if a particular tool allows it), but a narrower gene list will avoid noise and false positives which could hide the hunted gene.

As stated before, a common scenario consists of a research group in a wet lab hunting a gene involved in a particular disease. The traditional techniques used in the laboratory to identify the putative gene vary but they have in common that their output is not usually limited to one or more genes but to several megabases.

Validating all these genes one by one seems a too expensive and time-consuming task. Instead, using these regions as candidate sets in gene prioritization experiments can narrow down the list of candidates to an affordable number to manually validate.

The most common wet lab strategies applied when a gene of interest for a particular disease is to be found are:

- Genetic linkage.- This technique locates the gene of interest in a region neighbouring a piece of known DNA called a genetic marker and that is related to the disease. It has been used along with positional cloning in the last decades of the 20th century to find gene mutations that lead to monogenic disease such as cystic fibrosis and Huntington's disease [50, 51].
- Genome-wide association studies.- GWAS scans the full genome of a set of individuals with the aim of finding genetic alterations connected with a phenotype of interest. This high throughput approach has been used in the last years to identify a large number of robust associations between specific chromosomal loci and complex human diseases, such as type 2 diabetes and rheumatoid arthritis [52]
- Differential expression of genes in a disease tissue.- A microarray can also be used to select an initial candidate set. In this case, the list of candidate genes will not be a region but a group of genes that are differentially expressed among patients with respect to control cases.
- Chromosomal aberration.-If a chromosomal region is duplicated or deleted in a group of patients sharing a genetic condition, it becomes immediately a region of interest in a gene prioritization experiment.

1.3.2 TRAINING SET

The selection of the training genes by the user usually includes those genes that have already been linked to the disease. While the selection of the candidate set is basically based on the output of a wet lab experiment, the choice of the training set depends much more on the expertise of the user. He must take the decision whether adding or not to the training set a gene weakly linked to the disease, or a promising candidate still to be confirmed. Furthermore, a *wrong* candidate set, as long as it contains the gene of interest, would mean a larger ranking but, if the approach is good, the hunted gene would still rank high. However, a wrong training set would imply a wrong profile and the comparisons between candidate genes and the profile would lead to an inaccurate ranking.

A particularly difficult case arises with complex syndromes where different diseases classify under a single name. This is the case of e.g. leukemia, a type of cancer affecting blood and bone marrow that presents an uncontrolled proliferation of undifferentiated bone marrow cells. Under the name *leukemia*, different syndromes spread:

- Lymphoblastic or lymphocytic leukemia.- In this type of leukemia, cells created in the bone marrow that should develop into lymphocytes are abnormal, do not develop properly and grow quickly.
- Myeloid or myelogenous leukemia.- In this case, the abnormal cells are supposed to develop into red blood cells, some types of white blood cells and platelets.

These two types of leukemia can be further split into chronic and acute syndromes and in addition, there are other more rare types of the disease.

In chapter 5, we describe a method to relieve the user from the burden of choosing the right training set. We argue that the use of cluster analysis can help the user to obtain homogeneous training sets that will in general lead to more accurate rankings.

1.4 CLUSTER ANALYSIS

Clustering, or cluster analysis, corresponds to the assignment of objects to groups called clusters, in such a way that elements falling in the same cluster will be similar among them and dissimilar from the elements placed in other clusters. The number and the characteristics of the clusters are not known beforehand and for this reason, cluster analysis is also known as unsupervised learning. Clustering is widely used in many fields such as statistical pattern recognition [53], data mining [54], machine learning [55], information retrieval [56] or bioinformatics [57] and it has plenty of daily applications like image segmentation [58], target marketing [59], network intrusion detection [60] or financial fraud detection [61].

The evolution of clustering goes beyond 50 years of research and has consisted of an interdisciplinary effort made by taxonomists, social scientists, psychologists, biologists, statisticians, mathematicians, engineers, computer scientists, medical researchers and many others who collected and analyzed vast amounts of real data [62].

The previous definition of what clustering is implies the existence of the notion of similarity between objects of a database, in our case genes or gene products, so that they can be compared and assigned to the same or to different clusters. Defining this similarity is the keystone in a clustering algorithm, since the whole classification will depend on it. However, the nature of the elements to be clustered can be very diverse and the measure of similarity chosen will be one or another according to the type of attributes that these elements will contain.

1.4.1 TYPES OF DATA

A database consists of a set of records or *tuples* defined over a set of *attributes* which can take different *values*. A thorough taxonomy of the types of attributes can be found in [63]. Briefly, they can be classified either according to the size of the domain of the values that the attribute can take or based on the attributes measurement scales.

The size of the domain of the attributes indicates how many different values the attribute can take. On the one hand, an attribute will be called *continuous* if, between any two values, there exists an infinite number of other values. For instance, the height or the weight of a person are continuous values. On the other hand, an attribute will be discrete if the values that it can take are finite. For instance, the salary or the number of children of a person are discrete values. A special case of discrete attributes are those which can only take two values. These are called *binary* attributes. For instance, the gender of a person is a binary attribute.

The second way of classifying attributes is based on the existence of a measuring scale which allows values to be ordered and, therefore, give us the opportunity of comparing them. Suppose an attribute *i,* and two objects *x* and *z*, with values x_i and y_i for this attribute, respectively. Then, there are four different situations [64]:

1. If we can only distinguish values without being able to order them, i.e. we have either $x_i = y_i$ *or* $x_i \mathrel{!}= y_i$, we will say that the scale is nominal. *Nominal-scaled* values cannot be ordered. For instance, the birthplace of a person is a nominal-scale attribute.
2. *Ordinal-scaled* attributes are like nominal-scaled but with the possibility of being ordered. However, differences about values cannot be quantified. Therefore, the distinguished situations would be $x_i = y_i$, $x_i < y_i$ and $x_i > y_i$. For example, the Mohs scale of mineral hardness, which classifies minerals according to the ability of a harder material to scratch a softer one, but gives no clue about “how hard” a mineral is, is an example of ordinal-scaled attributes.
3. *Interval-scaled* values can be measured in a linear scale. They can not only be ordered, like *ordinal-scaled*, but the difference between them can be quantified. Examples include the Celsius scale of temperature.
4. *Ratio-scaled* values include interval-scaled and add a meaningful zero point. An example is the Kelvin temperature scale. The zero point of the Celsius scale of temperature is arbitrary and even negative numbers are used, but in the case of

Kelvin temperature scale, the zero point, the absolute zero, is not arbitrary but physically significant.

The first two types, nominal- and ordinal-scaled attributes, are known as *qualitative* or *categorical* attributes, while interval- and ratio-scaled attributes are called *quantitative* or *numerical*.

1.4.2 TRADITIONAL CLUSTERING APPROACHES

Originally, clustering methods were divided in two types: partitional and hierarchical methods. Since no *optimal* method was found after years of research, new perspectives were explored and nowadays we can distinguish up to five different strategies to deal with the clustering problem:

- Partitional clustering.- Given a database of *n* objects to be clustered, a partitional clustering algorithm constructs simultaneously *k* partitions of the data, optimizing an objective function. Enumerating all possible data groups and finding the optimum one is computationally infeasible (even for a small number of items, the number of possible partitions to be explored is huge). Therefore, most techniques using this approach start with an initial partition and then an iterative optimization, mostly a greedy algorithm, leads to its improvement. Examples of these techniques include K-means [65, 66] and its improved K-medoids methods, like PAM [67], CLARA [67] and CLARANS [68].
- Hierarchical clustering.- Given a database of *n* objects to be clustered, a hierarchical clustering algorithm constructs a tree of clusters called *dendrogram*. This family of algorithms can be further divided in *agglomerative* techniques, when the hierarchical decomposition is created in a bottom-up manner and *divisive*, when the tree is formed in a top-down fashion.
    - Agglomerative clustering algorithms follow a bottom-up strategy. The initial solution is a set of as many clusters as objects to be clustered, and iteration after iteration, the most similar clusters are merged until a unique cluster including all the objects is reached.
    - Divisive clustering algorithms follow a top-down strategy. The initial scenario is a unique cluster containing all the items to be clustered. In every iteration, the most dissimilar clusters are divided until there are as many clusters as items.

  Examples of these algorithms include CURE [69] and BIRCH [70].
- Density-Based clustering.- These methods explore the data space searching for dense regions of items separated by regions of low density. Normally, the cluster grows as long as the number of objects in the neighborhood is greater than a certain parameter. Examples of these techniques include DBSCAN[71], OPTICS[72] and DENCLUE [73].

- Grid-based clustering.- The data space is divided in a finite number of cells. Clusters will be formed from adjacent groups of cells whose density of objects exceeds a certain threshold. STING [74], WaveCluster [75] and CLIQUE [76] are well known examples.
- Model-Based clustering.- These methods search for the optimum parameters that make a mathematical model fit with the data to be clustered. They are mostly based on the assumption that the data are generated by a mixture of probability distributions. The most known example is the *Expectation-Maximization* algorithm [77].

1.4.3 CATEGORICAL CLUSTERING

Originally, most of clustering algorithms developed focused on numerical data**.** This did not happen due to the relative amount of numerical data over categorical data in databases but because of the ease of developing a clustering algorithm based on numerical data and, therefore, on the concept of distance.

However, the important presence of categorical data during the massive growth of data storage during the last years, made the clustering community to turn their eyes to this new kind of clustering problem where the difference among objects must be measured using other strategies than distance.

The first strategy to cope this problem was the so called *conceptual clustering* where objects were classified into clusters that represented certain descriptive concepts rather than into clusters defined by distance among the objects [78]. *Cobweb* [55] became very popular. This algorithm uses incremental learning to build a dendrogram, but instead of using divisive or agglomerative strategies, it dynamically builds it by processing a single data point at a time. Every node of the tree represents a certain concept which is associated with a set of objects, each of one described by a list of Boolean properties. In every new iteration, any dataset point to be clustered will compare its properties with the total addition of all the different properties of the points in each cluster and, according to a user-dependent parameter will join a cluster or will seed a new one.

Some of the first categorical clustering methods tried to adapt strategies that were used in numerical clustering to categorical data. This is the case of *k-modes* and its successors. *K-modes* [79] is an extension of the classic clustering algorithm k-means adapted to categorical data. Like in k-means, a partitioning algorithm, k-modes partitions a data set into a given number of clusters such that an objective function is optimized. The differences with the original algorithm concern the distance measure used (based on matches and mismatches of items instead of using Euclidean distance), the use of *modes* instead of *means* and the use of a frequency based method to update modes in every iteration. K-modes has been improved using weighted attributes based on the ratio of frequency of attribute values in the data set [80]

Other modifications of the k-means algorithm is *k-histograms* [81] which replaces the means of the clusters with histograms which are dynamically updated during the clustering process.

There exist also methods which tackle the clustering problem as a search in a graph. *ROCK* [82] is a hierarchical clustering algorithm which considers two *records* to be *neighbors* when the distance between them, based on the Jaccard coefficient, is lower than a pre-specified threshold. A *link* connects two records if they share a neighbor. The criterion that ROCK follows is to group records according to the number of links connecting them. Unfortunately, since this is an agglomerative algorithm, is not applicable to large datasets. Some modifications have been proposed that outperform the original ROCK algorithm: *VBACC* [83] assumes that, if the data set has high dimensionality, the pre-specified threshold can be omitted since it leads to a too sparse graph and the clustering algorithm fails to properly partition; *QROCK* [84] sees the dataset points as vertices of a graph and finds the clusters by determining the connected components of the graph and *QNNS* [85] can select *qualified neighbors* automatically without pre-specifying the threshold parameter.

Other graph-based clustering methods are *Click* [86]**,** a clustering algorithm which finds clusters based on a search method for k-partite maximal cliques and *STIRR* [87], which undertakes clustering as a partitioning problem in a hypergraph and solves it using non-linear dynamical systems.

Some other methods use a summary approach. These methods do not store all the information of the tuples to be clustered, but use a summary, speeding up the clustering. Two examples are *CACTUS* [88] and *LIMBO* [89]. CACTUS assumes that the domain sizes of categorical attributes are small. It uses two types of summaries, intra and inter-cluster, and works in three phases: summarization, clustering and validation. LIMBO is a scalable and hierarchical clustering algorithm based on the idea that keeping a summary that describes tuples and clusters is sufficient. The clustering process uses a notion of distance inspired in the *Information Bottleneck* framework [90].

Another categorical clustering method, using a different approach is *COOLCAT* [91], which is based on the entropy of the clusters. This algorithm takes advantage of the fact that the entropy of a set is inversely proportional to the amount of common attribute values of the set. In a first phase, the algorithm defines a suitable number of clusters and in a second phase it allocates the rest of the points to the clusters minimizing the value of the entropy of the overall cluster.

And finally, there are also methods which cluster tuples according to the frequency of the values of attributes. *Squeezer* [92] is a scalable and incremental clustering algorithm based on the maximization of the *support* of the attribute values. This measure is the number of tuples in a cluster containing the value. In only one scan over the dataset, every item either seeds a new cluster or is grouped in an existing one, depending on the pre-specified similarity threshold. Being a one-scan-algorithm, the order in which the dataset points are inputted influences the partition of the data.

The output of most of the clustering algorithms, not only the methods addressed to categorical data but in general, depends on the user, since initial parameters introduced by the user are necessary. Not all the methods require them and the ones which do, do not always need the same type of input. However, there is one parameter common to many tools: the number of clusters.

The presence of these parameters, along with the strong dependency of the cluster algorithm on the type of data, makes very difficult to quantitatively compare clustering tools. A fair comparison will only be done if exhaustive experiments covering all possible problems are undertaken. This has not been done, so far, but partial work, compares a selection of specific tools in specific problems. Most of cases where comparisons among tools have been explicitly undertaken have used the UCI machine learning repository datasets (http://www.cis.uci.edu/~mlearn/MLRepository.html), using specially mushroom, zoo and congressional voting datasets[81, 84–86, 88, 89, 92] but others have created synthetic databases based on *ad-hoc* generators [87]. However, these publications can show different results when comparing the same tools on the same data. Even if the database is the same, there is a non negligible human factor which can bias the results. Since not all the tools need the same type of input, making a fair comparison among tools becomes even more difficult. In addition, some comparisons are not thoroughly done and use different software platforms, with different technical characteristics.

1.4.4 TRANSACTIONAL CLUSTERING

Within the categorical data, we can distinguish a particular type of objects, called *transactions*. A transaction is a set of related items which can be viewed in a database as records with attributes, each corresponding to a single item**.** An example of transactional data is the market basket data, for instance t_1={beer, milk}, t_2*={milk, oranges}* and it can be represented by t_1={1,1,0} and t_2={0,1,1} if the first attribute stands for beer, the second for milk and the third one for oranges. The volume of transactional databases is normally very large, either in number of records and in the amount of attributes and, eventually, dimensions in the data space. These two common characteristics in transactional data lead to a, usually, very sparse database. For example, in a market basket database, a very low percentage of customers will have any product in common with each other. The size of the database, the high dimensionality and hence the sparsity make either traditional numerical and categorical clustering techniques unsuitable. New algorithms have been developed in the last years to cope with this problem. These new strategies define global criterion functions instead of local ones, like some of the categorical clustering algorithms introduced in the previous sections do [82, 87, 88]. The difference between both types of criterion functions is that the latter works with a pairwise similarity between data points while the former computes at cluster level. That makes the locally defined criterion function based algorithms unsuitable for large databases as the computation cost grows very fast.

But not only new algorithms have been designed from scratch, since also some categorical algorithms have been adapted to the characteristics of the transactional databases, like the previously described VBACC as an improvement of ROCK.

The first algorithm specifically designed for transactional data was *Largeitem* [93]. This clustering algorithm is based on the notion of *large item*. A transaction item *i* will be defined as a large item in a cluster C_i if, for a pre-specified *support* θ (between 0 and 1), the number of transactions in the cluster is at least $\theta^*|C_i|$. Otherwise, the transaction item will be defined as a small item. The clustering algorithm using large items tries to minimize the *cost* of the solution, minimizing the number of small items and maximizing the number of large items grouped in the same cluster.

Other approaches are *OAK* [94], an agglomerative hierarchical clustering algorithm that uses cosine similarity as a distance measure between transactions, *CLOPE* [95], which clusters based on the maximization of histograms of every cluster and *k-todes* [96] which uses *category-based adherence*, based on the average distance of the items in a transaction to a cluster, as a global measure to be minimized..

1.4.5 CONCLUSION

Cluster analysis has become very important during the last years due to the massive growth of databases following the dramatic increase on computing power. Mining documents on Internet [97] and finding genes with similar functions in microarrays [98]are two of many applications of these techniques.

However, clustering is a very difficult problem. The definition of what a cluster is, is rather vague. One has to decide on an appropriate similarity measure to compare objects as well as on an objective function to drive the process. Both choices offer possibilities to express domain knowledge. While this flexibility increases the applicability, an objective comparison of clustering approaches over a set of different problems is made hard if not impossible. Clustering also depends deeply on the nature of the data to be clustered. The type of the attributes is crucial, but also the sparsity and the number of attributes of every object have to be taken into account before choosing the right method. This makes possible that using different algorithms on the same data, the results can be completely different [62]. There is not a gold standard, and for a particular problem, it can happen that an old and classic method works better than a recently proposed one. In fact, K-means, proposed over 50 years ago is still widely used and, moreover, has been taken as a standard to be improved when categorical and transactional clustering problems have been tackled.

A particularly difficult type of cluster analysis, involves categorical data. In this type of problem, distance cannot be used to measure similarity or dissimilarity among objects and other strategies have to be selected. Specific algorithms have been proposed particularly in the last years to cope with this problem and this section lists them so that a more clear vision can be taken of this research field.

When the records to be clustered are transactional items, that is to say, a set of related items with attributes corresponding to a single item, then we are before a transactional cluster analysis problem. Numerical clustering algorithms are useless to tackle this

problem due to the curse of dimensionality, a situation that appears when the number of attributes of each object (and hence, dimensions) becomes too high. When the number of dimensions increases, the distance from a point to the closest one approaches the distance between this point and the furthest one [99]. Some categorical clustering algorithms also fail facing this type of data because of the selection of a local criterion function to optimize during the process of clustering, which is too time and resource consuming when the database is large.

1.5 AIMS AND OBJECTIVES

The gene prioritization problem was defined a decade ago and since then, many different computational solutions have been created to tackle it. After years of research, the set of gene prioritization tools is large and heterogeneous and include approaches based on different strategies, such as training sets of genes or keywords describing the knowledge of the disease, different ways to input the candidate set to be ranked, multiple types of data sources and diverse types of outputs.

The objective of this work is double. First of all, we intend to clarify the fast evolving gene prioritization field by thoroughly describing and comparing the gene prioritization approaches available as free online services. We point out the similarities and the differences, stressing in which cases a particular tool or group of tools is more suitable than others and comparing a representative subset of them in terms of performance.

And second, we aim at increasing the performance of gene prioritization experiments by a first time statistical supported gene prioritization tools integration and by a cluster analysis based preprocessing step for training set based gene prioritization tools. Both cases are validated and applied to real biological data.

1.6 STRUCTURE OF THE THESIS AND PERSONAL CONTRIBUTION

In this thesis, we have given an insight into gene prioritization. We have thoroughly described the current gene prioritization tools freely available as web interfaces, we have compared some of them in terms of performance and reliability and we have combined the top performing ones in a new strategy to propose new meaningful candidates for a number of congenital and complex diseases. Furthermore, we propose a cluster analysis preprocessing step for training set based gene prioritization tools in order to obtain better rankings.

Chapters 2 and 3 depict the state of the art in gene prioritization. Chapter 2 reviews all the gene prioritization tools that have been developed in the last years and that offer a free web based interface. Chapter 3 goes one step beyond in the review of gene prioritization

tools and sets up a benchmark of 42 diseases where the performance of the tools is compared.

Chapter 4 presents a study where the highest quality gene prioritization tools based in the benchmark previously described are combined in a novel two-layer based strategy and their prioritization power is applied to three different genetic conditions.

Chapter 5 introduces the use of cluster analysis as a preprocessing step in gene prioritization in order to obtain homogeneous training sets with the aim of producing higher quality rankings.

Chapters 2, 3 and 4 have been jointly produced by the Ph.D. candidate and the mentioned co-authors in terms of the conception of the idea and development of the study. In chapter 2, the Ph.D. candidate has set up the initial idea, reviewed one third of the gene prioritization tools, has mined the literature in search of tool validations and has written the manuscript. In chapter 3, the Ph.D candidate has devised the study, has performed experiments with three tools and up to five different configurations and has written the paper. In chapter 4, the Ph.D. candidate has initiated the approach, has run two gene prioritization tools for three diseases, has collaborated in the interpretation of the final ranking related to asthma and has written the manuscript.

In chapter 5, the Ph.D. candidate has set up the initial idea, devised the strategy, implemented the cluster analysis, analyzed the results and finally has written the manuscript.

Chapter 2
Summary

The gene prioritization problem has attracted a deep attention from the bioinformatics community during the last decade, since the problem was defined in 2002 [17]. Since then, many different approaches have been developed to tackle this problem. The existence of so many different tools with similar objective made the choice of one of them a difficult decision for a potential user, what is just the contrary of the main purpose of gene prioritization tools: to ease the work of the wet-lab researchers.
Our main aim in this publication has been to make the decision of which gene prioritization tool allows the easier use for any potential user, regardless of his expertise with computers.
We exhaustively review the class of freely accessible gene prioritization solutions offering a web interface. Therefore, we have discarded gene prioritization tools without a web interface to allow users, who are not experts in the use of the computers, could make the most of this paper.
We point out the tool differences in terms of input, output and databases used. Input in gene prioritization can vary in terms of both training genes and candidate genes. The training data can consist of known genes and/or keywords and the candidate data can include a chromosomal region, a list of genes differentially expressed or even the full genome. Gene prioritization tools also differ in the output that they offer. Whereas most of them return a ranking of genes, a selection of promising candidates is also a valid output. In both cases, a statistic support of the results can be provided. As for the databases, gene prioritization tools can work with functional annotation data, protein-protein interaction, text mining, pathway data, expression values, sequence information, phenotypic data and others. Therefore, with so many variants and as stated before, selecting the right tool could be a difficult step for the user and our main aim has been to make this decision easier.
Based on these variants, we also propose in this chapter a decision tree that puts the different possibilities in a visual way and that can be used by the final user to find the most appropriate tool for his needs.
Furthermore, we include a bibliography with the validations of every tool and different publications where performance comparisons among tools can be found.
Finally, we have developed a website containing up-to-date information about these tools (www.esat.kuleuven.be/gpp) In the last months, new tools have been added showing that the gene prioritization portal can be a reference site for researchers interested in the gene prioritization problem [100–102].

This chapter has been published in 2010 as an electronic version and in 2011 as a printed version in the *Briefings in Bioinformatics* journal:

Tranchevent, L.-C., Capdevila, F.B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2011). *A guide to web tools to prioritize candidate genes. Brief. Bioinformatics 12*, 22–32.

Personal contribution

This chapter has been jointly produced by the Ph.D. candidate and the mentioned co-authors in terms of the conception of the idea and development of the study. In particular, the Ph.D. candidate has set up the initial idea, reviewed one third of the gene prioritization tools, has mined the literature in search of tool validations and has written the manuscript.

BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 1. 22–32
Advance Access published on 21 March 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@oup.com

Downloaded from http://bib.oxfordjournals.org/ at KU Leuven - Faculteit Rechtsgeleerdheid on September 12, 2012

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 - Faculteit Rechtsgeleerdheid on September 12, 2012

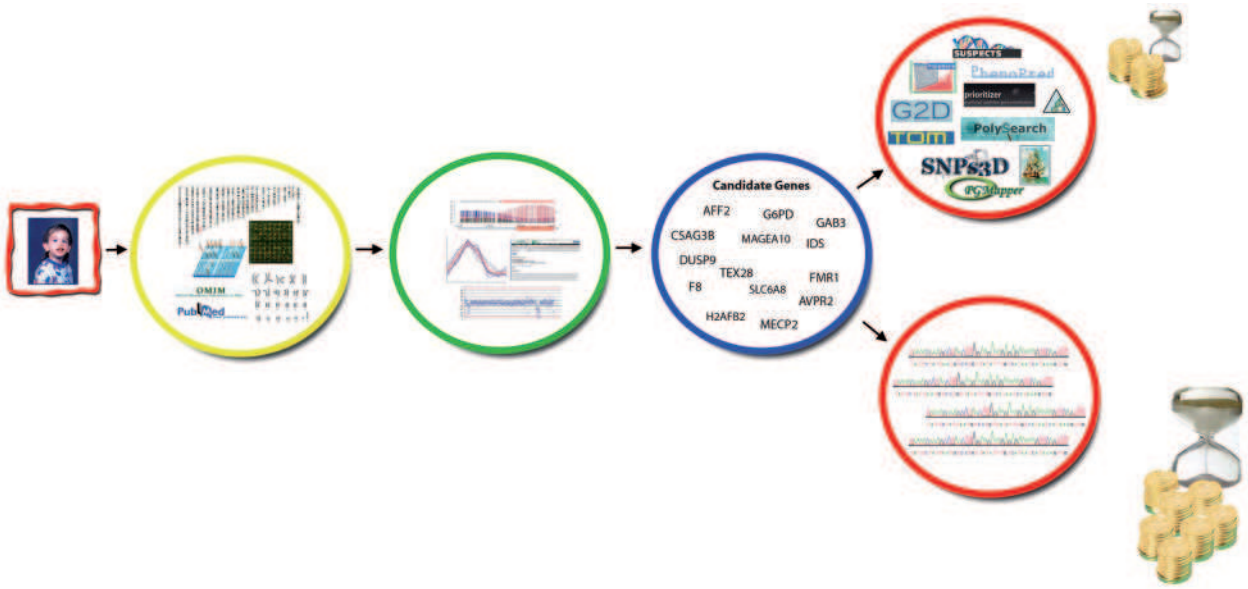


Figure 1: A major challenge in human genetics is to unravel the genetic variants and the molecular basis that underlay genetic disorders. In the past decades, geneticists have mainly used high-throughput technologies (such as linkage analysis and association studies). These technologies usually associate 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.

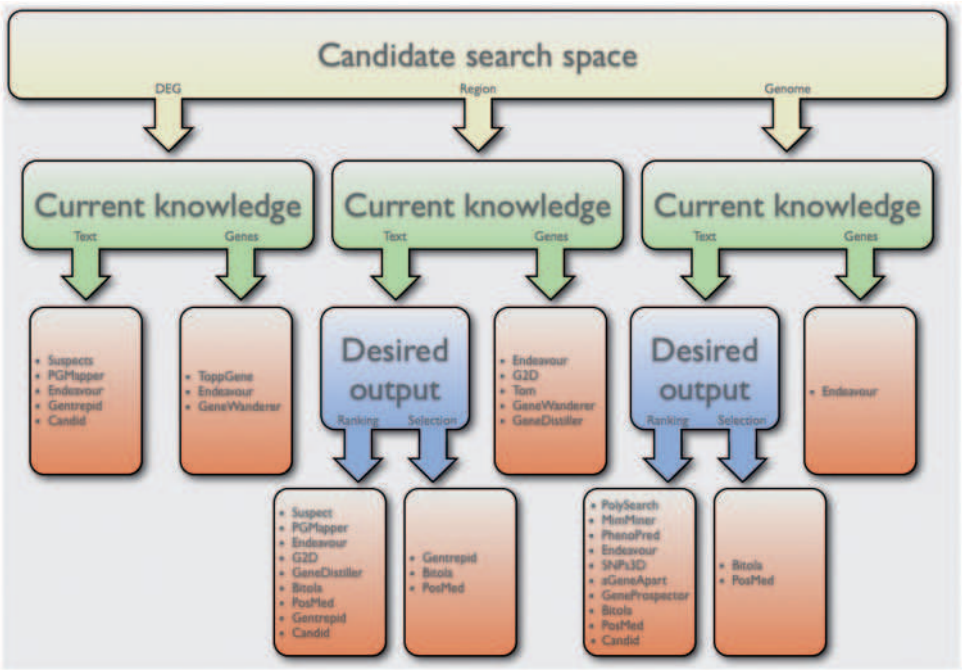


Figure 2: Decision tree that categorizes the 19 gene prioritization tools according to the outputs they use and the outputs they produce. This tree is designed to support the end users in their decision so that they can choose the tools that suit best their needs. By starting from the first question 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 - Faculteit Rechtsgeleerdheid on September 12, 2012

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 - Faculteit Rechtsgeleerdheid on September 12, 2012

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-β, 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 - Faculteit Rechtsgeleerdheid on September 12, 2012

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 - Faculteit Rechtsgeleerdheid on September 12, 2012

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.

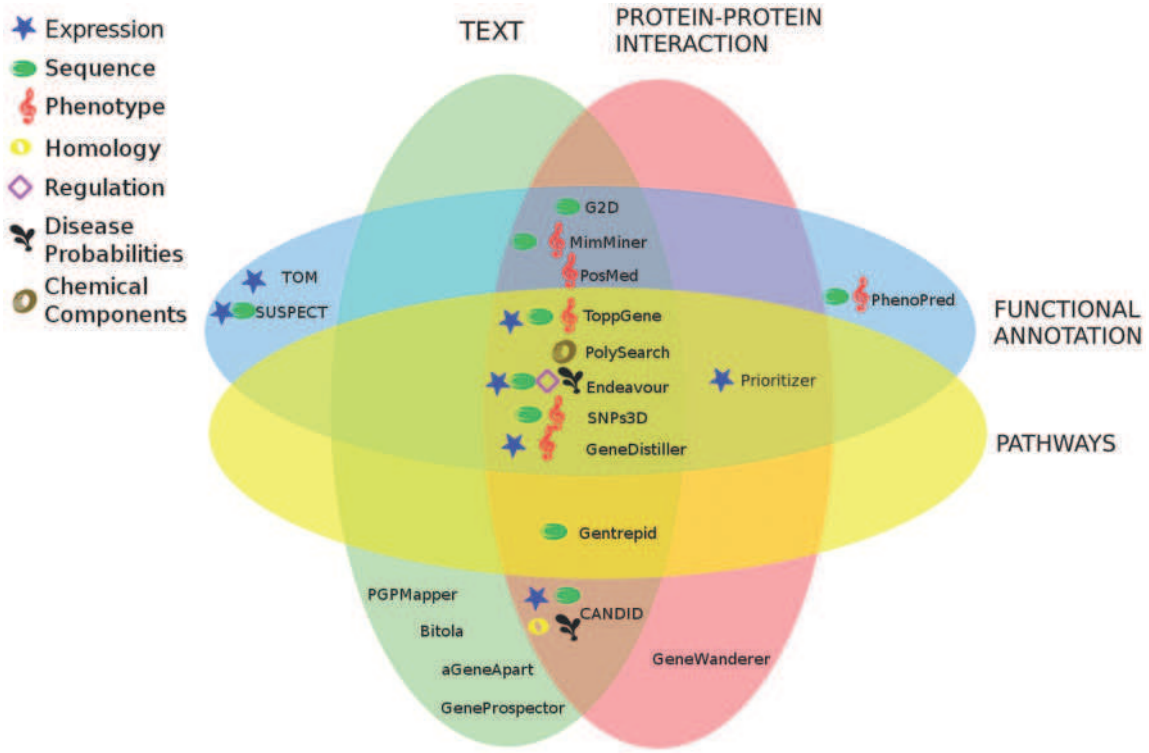


Figure 3: Repartition of the 19 tools according to the data sources they use. The four data sources most commonly used are Text (functional and interactions mining), protein–protein interactions, functional annotations and pathways and are therefore represented as large ellipses. The additional seven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

Downloaded from http://bib.oxfordjournals.org/ at KU Leuven - Faculteit Rechtsgeleerdheid on September 12, 2012

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 - Faculteit Rechtsgeleerdheid on September 12, 2012

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 - Faculteit Rechtsgeleerdheid on September 12, 2012

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 - Faculteit Rechtsgeleerdheid on September 12, 2012

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 - Faculteit Rechtsgeleerdheid on September 12, 2012

Supplementary material for the manuscript entitled “A guide to web tools to prioritize candidate genes” by Tranchevent *et al.*

	Text (coocurrance)	**Text (functional mining)**	**PPIs**	**Functional Annotations**	**Pathways**	**Expression**	**Sequence**	**Phenotype**	**Conservation/ Homology**	**Regulation**	**Disease probabilities/ scores**	**Chemical Components**
SUSPECT				x		x	x					
ToppGene	x		x	x	x	x	x	x				
PolySearch		x	x	x	x							x
MimMiner		x	x	x			x	x				
PhenoPred			x	x			x	x				
PGMapper		x										
Endeavour		x	x	x	x	o x	x			x	o x	
G2D	x		x	x			x					
TOM				x		x						
SNPs3D		x	x	x	x		x	x				
GenTrepid		x	x		x		x					
GeneWanderer			x									
Bitola		x										
CANDID		x	x			x	x		x		o	
aGeneApart		x										
GeneProspector		x										
PosMed	x	x	x	x				x				x
GeneDistiller	x		x	x	x	x		x				

Supplementary Table 1: Data sources used by the 19 tools. We have defined 12 distinct types, each type can correspond to several databases (e.g., Reactome and Kegg are two pathway databases). A cross means that a data source is available for one tool. A circle means that the user can add its own data source of that type.

Tool	Reference	Compared with
SUSPECTS	Teber *et al.* (2009) [1]	GenTrepid, G2D
	Tiffin *et al.* (2006) [2]	G2D, GenTrepid
	Thornblad *et al.* (2007) [3]	PosMed
	Huang *et al.* (2008) [4]	Endeavour
ToppGene		
PolySearch		
MimMiner		
PhenoPred		
PGMapper		
Endeavour	Hutz *et al.* (2008) [5]	CANDID
	Köhler *et al.* (2008) [6]	GeneWanderer
	Huang *et al.* (2008) [4]	Suspects
G2D	Teber *et al.* (2009) [1]	GenTrepid, SUSPECTS
	Tiffin *et al.* (2006) [2]	SUSPECT, GenTrepid
TOM		
SNPs3D		
GenTrepid	Teber *et al.* (2009) [1]	G2D, SUSPECT
	Tiffin *et al.* (2006) [2]	G2D, SUSPECT
GeneWanderer	Köhler *et al.* (2008) [6]	Endeavour
Bitola		
CANDID	Hutz *et al.* (2008) [5]	Endeavour
aGeneApart		
GeneProspector		
PosMed	Thornblad *et al.* (2007) [3]	SUSPECTS

Supplementary Table 2: Collection of studies that compare different gene prioritization tools by applying them either to real biological problems or to a common benchmark.

Tool	Reference	Disease
SUSPECTS	Tiffin *et al.* (2006) [2]	Type 2 Diabetes
	Elbers *et al.* (2007) [7]	Type 2 Diabetes
	Teber *et al.* (2009) [1]	Type 2 Diabetes
	Huang *et al.* (2008) [4]	Osteoporosis
ToppGene	Sinha *et al.* (2008) [8]	p53-mediated tumorigenesis
PolySearch		
MimMiner		
PhenoPred		
PGMapper	Xiong *et al.* (2008) [9]	Arthritis
Endeavour	Huang *et al.* (2008) [4]	Osteoporosis
	Windelinckx *et al.* (2007) [10]	Muscle strength
	Elbers *et al.* (2007) [7]	Common obesity and type 2 Diabetes
	Cheung *et al.* (2008) [11]	Bone mineral density variation and fracture risk association
	Liu *et al.* (2008) [12]	BMD and bone structure
	Tzouvelikis *et al.* (2007) [13]	Pulmonary fibrosis
	Osoegawa *et al.* (2008) [14]	Cleft lip and palate
	Adachi *et al.* (2007) [15]	Adipocyte proteome
	Vanden Bempt *et al.* (2007) [16]	Breast cancer
	Storey *et al.* (2009) [17]	Spinocerebellar ataxia
	Aerts *et al.* (2009) [18]	Atonal mediated neural development
	Katsanou *et al.* (2009) [19]	Placental Branching Morphogenesis and Embryonic Development
	Sookoian *et al.* (2009) [20]	Type 2 diabetes
G2D	Teber *et al.* (2009) [1]	Type 2 Diabetes
	Tiffin *et al.* (2006) [2]	Type 2 Diabetes
	Tiffin *et al.* (2008) [21]	Metabolic syndrome
	Elbers *et al.* (2007) [7]	Common obesity and type 2 Diabetes
TOM		
SNPs3D		
GenTrepid	Teber *et al.* (2009) [1]	Type 2 Diabetes
	Sparrow *et al.* (2008) [22]	Spondylocostal dysostosis (SCD)

GeneWanderer		
Bitola		
CANDID		
aGeneApart	Pasmant *et al.* (2008) [23]	NF1 contiguous gene syndrome
GeneProspector		
PosMed		
GeneDistiller		

Supplementary Table 3: Collection of the biological applications found in the literature that use one of the tools of the present review. Redundancy with the Suppl. Table 2 is possible when several tools were used in conjunction.

References

1. 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
2. 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; 34:3067-3081
3. Thornblad, TA, Elliott, KS, Jowett, J et al. Prioritization of positional candidate genes using multiple web-based software tools. Twin Res Hum Genet 2007; 10:861-870
4. Huang, Q, Li, GHY, Cheung, WMW et al. Prediction of osteoporosis candidate genes by computational disease-gene identification strategy. J. Hum. Genet 2008; 53:644-655
5. Hutz, JE, Kraja, AT, McLeod, HL et al. CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genet. Epidemiol 2008; 32:779-790
6. Köhler, S, Bauer, S, Horn, D et al. Walking the interactome for prioritization of candidate disease genes. Am. J. Hum. Genet 2008; 82:949-958
7. 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
8. Sinha, AU, Kaimal, V, Chen, J et al. Dissecting microregulation of a master regulatory network. BMC Genomics 2008; 9:88
9. Xiong, Q, Qiu, Y Gu, W. PGMapper: a web-based tool linking phenotype to genes. Bioinformatics 2008; 24:1011-1013
10. Windelinckx, A, Vlietinck, R, Aerssens, J et al. Selection of genes and single nucleotide polymorphisms for fine mapping starting from a broad linkage region. Twin Res Hum Genet 2007; 10:871-885
11. Cheung, C, Sham, PC, Chan, V et al. Identification of LTBP2 on chromosome 14q as a novel candidate gene for bone mineral density variation and fracture risk association. J. Clin. Endocrinol. Metab 2008; 93:4448-4455
12. Liu, X, Liu, Y, Liu, J et al. A bivariate whole genome linkage study identified genomic regions influencing both BMD and bone structure. J. Bone Miner. Res 2008; 23:1806-1814
13. Tzouvelekis, A, Harokopos, V, Paparountas, T et al. Comparative expression profiling in pulmonary fibrosis suggests a role of hypoxia-inducible factor-1alpha in disease pathogenesis. Am. J. Respir. Crit. Care Med 2007; 176:1108-1119
14. Osoegawa, K, Vessere, GM, Utami, KH et al. Identification of novel candidate genes associated with cleft lip and palate using array comparative genomic hybridisation. J. Med. Genet 2008; 45:81-86
15. Adachi, J, Kumar, C, Zhang, Y et al. In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. Mol. Cell Proteomics 2007; 6:1257-1273
16. Vanden Bempt, I, Drijkoningen, M De Wolf-Peeters, C. The complexity of genotypic alterations underlying HER2-positive breast cancer: an explanation for its clinical heterogeneity. Curr Opin Oncol 2007; 19:552-557
17. Storey, E, Bahlo, M, Fahey, M et al. A new dominantly inherited pure cerebellar

ataxia, SCA 30. J. Neurol. Neurosurg. Psychiatr 2009; 80:408-411
18. Aerts, S, Vilain, S, Hu, S et al. Integrating computational biology and forward genetics in Drosophila. PLoS Genet 2009; 5:e1000351
19. Katsanou, V, Milatos, S, Yiakouvaki, A et al. The RNA-binding protein Elavl1/HuR is essential for placental branching morphogenesis and embryonic development. Mol. Cell. Biol 2009; 29:2762-2776
20. Sookoian, S, Gianotti, TF, Schuman, M et al. Gene prioritization based on biological plausibility over genome wide association studies renders new loci associated with type 2 diabetes. Genet. Med 2009; 11:338-343
21. 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
22. Sparrow, DB, Guillén-Navarro, E, Fatkin, D et al. Mutation of Hairy-and-Enhancer-of-Split-7 in humans causes spondylocostal dysostosis. Hum. Mol. Genet 2008; 17:3761-3766
23. Pasmant, E, de Saint-Trivier, A, Laurendeau, I et al. Characterization of a 7.6-Mb germline deletion encompassing the NF1 locus and about a hundred genes in an NF1 contiguous gene syndrome patient. Eur. J. Hum. Genet 2008; 16:1459-1466

Chapter 3

Summary

This chapter expands the review of the previous chapter adding a quantitative component following the spirit of CASP for protein structure prediction.

In our effort to help the gene prioritization final user to select the most convenient tool for his purposes, we have compared several web based and freely accessible gene prioritization tools in terms of performance and reliability. We have measured the tool performance based on the position in the final ranking of a set of recently discovered disease genes. As for reliability, we have taken into account the percentage of experiments where the validating gene was returned by the tool as part of the final ranking.

The main differences between this approach and other similar work is that we have used recently published disease genes as validating genes. During six months, we have selected from scientific literature those genes published as linked to a particular genetic condition. We have run our experiments in the 48 hours following to the publication to ensure that the databases used by the tools where still not updated with this information.

The input for all the gene prioritization tools has been the same and it has consisted of a set of keywords built by ourselves and a chromosomic region of about 10 Mb encompassing the validating gene.

In this way, we have run 43 times several gene prioritization tools and we have recorded both the final position of the validating gene and whether this gene was taken into account by the tool or not. Several statistical measures have been calculated.

Furthermore, we have integrated the tools and compared the combined ranking with the individual ones and the results show better figures in the combined approach, giving a strong support to the work that we present in chapter 4.

This chapter has been submitted to the *Bioinformatics* journal on July 2012.

Personal contribution

This chapter has been jointly produced by the Ph.D. candidate and the mentioned co-authors in terms of the conception of the idea and development of the study. In particular, the Ph.D candidate has devised the study, has performed experiments with three tools and up to five different configurations and has written the paper

Systems Biology

An unbiased evaluation of gene prioritization tools

Daniela Börnigen[1,2,*], Léon-Charles Tranchevent[1,*], Francisco Bonachela-Capdevila[3,*], Koenraad Devriendt[4], Bart de Moor[1], Patrick De Causmaecker[3], and Yves Moreau[1#]

[1]Department of Electrical Engineering, ESAT-SCD, IBBT-KULeuven Future Health Department, Katholieke Universiteit Leuven, Leuven, Belgium
[2]Biostatistics Department, Harvard School of Public Health, Harvard University, Boston, MA, USA
[3]CODeS Group, ITEC-IBBT-KULEUVEN, Katholieke Universiteit Leuven campus Kortrijk, Kortrijk, Belgium
[4]Center for Human Genetics, Katholieke Universiteit Leuven, Leuven, Belgium
*Contributed equally to this work

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Gene prioritization aims at identifying the most promising candidate genes among a large pool of candidates—so as to maximize the yield and biological relevance of further downstream validation experiments and functional studies. During the past few years, several gene prioritization tools have been defined and some of them have been implemented and made available through freely available web tools. In this study, we aim at comparing the predictive performance of eight publicly available prioritization tools on novel data. We have performed an analysis in which 42 recently reported disease gene associations from literature are used to benchmark these tools before the underlying databases are updated.

Results: Cross-validation on retrospective data provides performance estimate likely to be overoptimistic because some of the data sources are contaminated with knowledge from the disease-gene association. Our approach mimics a novel discovery more closely and thus provides more realistic performance estimates. There are however marked differences, and tools that rely on more advanced data integration schemes appear more powerful.

Contact: yves.moreau@esat.kuleuven.be

1 INTRODUCTION

A major challenge in human genetics is to discover novel disease causing genes, both for Mendelian and complex disorders. Identifying disease genes is a crucial first step in unraveling molecular networks underlying diseases, and thus understanding disease mechanisms, also towards the development of effective therapies. The discovery of a novel disease gene often starts with a cytogenetic study, a linkage analysis, a high-throughput omics experiment, or a genome-wide association study (GWAS). However, these studies do not always pinpoint the disease gene uniquely, but often result in large lists of candidate genes that are potentially relevant (Hardy and Singleton, 2009). Moreover, recent advances in next-generation sequencing offer promising opportunities to explore the genomic alterations of patients (Schuster, 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 resequencing, pathway or expression analysis, is still expensive and time consuming. An efficient way to reduce the validation cost is to narrow down the large list of candidate genes to a small and manageable set of highly promising genes, a process called 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 currently available.

In the past few years, many gene prioritization methods have been proposed, some of which have been implemented into publicly available tools that users can freely access and use (Moreau et al., 2012; Doncheva et al., 2012; Piro et al., 2012; Tiffin 2011, Oti 2011; Tranchevent et al., 2010). Information about these tools is summarized in our Gene Prioritization Portal (http://www.esat.kuleuven.be/gpp) that currently describes 33 prioritization tools. This web site 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, or when selecting candidates for a medium-throughput screen (instead of low-throughput validation). Another example is the study of a poorly characterized disorder, for which a prioritization tool not relying on a set of known disease genes might be more suited. Recently, several studies have demonstrated that gene prioritization tools can help geneticists to discover novel disease genes (Thienpont et al., 2010; Calvo et al., 2006). For instance, a KIF1A mutation was discovered in hereditary spastic paraparesis patients after KIF1A was predicted to be the best candidate gene from the locus using multiple prioritization tools (Erlich et al., 2011). Another study discovered homozygous mutations in the PTRF-CAVIN gene in patients with congenital generalized lipodystrophy with muscle rippling after PTRF-CAVIN was predicted as the most probable candidate gene for high expression

#To whom correspondence should be addressed.

© Oxford University Press 2005

in muscle and adipose tissue (Rajab et al., 2010). A third study identified the HHEX gene to be associated with type 2 diabetes in a Dutch cohort after investigating the T2D-susceptibility loci using candidate gene prioritization (Vliet-Ostaptchouk et al., 2008).

However, beyond these conceptual differences, one essential parameter to consider when selecting gene prioritization tools is their respective performance—that is, their ability to identify the true positive genes as promising candidate genes in order to maximize the yield of the follow-up experimental validation. 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 existing methods (Hutz et al., 2008; Köhler et al., 2008; Thornblad et al., 2007). However, these benchmarks are most of the time cross-validations of gold-standard disease data sets (e.g., known data). Therefore, the estimation of the performance is likely an overestimate of the real performance (i.e., on novel data). Because different types of data are dependent on each other (for example, GO annotation, KEGG pathway membership, and MEDLINE abstracts) it becomes impossible to remove all cross-talk effects between data sources (e.g., removing MEDLINE data does not remove all information from the biomedical literature since much of it is present in GO and KEGG) to prevent contamination of the prediction of the disease gene by actual retrospective knowledge of this association. This makes it challenging to create benchmarks on retrospective data that are indicative of the performance of the method in an actual research setting. Next to benchmarking, some studies use several prioritization methods to analyze disease associated loci, mostly for type 2 diabetes and obesity (Tiffin et al., 2006; Elbers et al., 2007; Teber et al., 2009). However, the results have not been experimentally validated, which means that it is not possible to identify which methods made better predictions. Also, a few studies combine computational and experimental analysis: in silico generated hypothesis are then validated in vivo. We have, for instance, performed a computationally-supported genetic screen in Drosophila that led to the identification of 12 novel atonal genetic interactors (Aerts et al., 2009). Although useful, such studies often rely on the use of a single tool and therefore cannot be used to compare different approaches. They also give no indication of the performance of the method in general, but only illustrate it on a single well-validated case.

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 selected recently reported disease gene associations from literature and use several gene prioritization tools to make predictions immediately after publication (typically within two days). 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, and therefore the estimation of the performance is more accurate. It has to be mentioned that we compare tools and not the underlying algorithms (we see a tool as an algorithm plus some data sources), because this is what is most relevant to geneticists.

2 METHODS

2.1 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 predict potential novel disease genes which can then be experimentally validated. We have therefore not selected the tools whose ranking strategies depend exclusively on text as they would most likely work only when the novel disease gene was already considered a good candidate gene prior to discovery. One exception is Candid that also uses other data sources beside MEDLINE (e.g., protein domains, interactions, and expression data). In total, we have selected eight 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), Endeavour (Aerts et al., 2006), and Pinta (Nitsch et al., 2010). The tools are run with their settings recommended by the developers. When applicable, multiple configurations are defined to explore several possibilities (for instance, several ranking algorithms within one tool). Originally, Pinta was developed to use expression data as input data, but here, we replace the continuous data (coming from expression data) with binary data using training genes: a *1* is inputed for each training gene, and a *0* is associated to the other genes. 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. Pinta and Endeavour support both genome-wide and candidate set based prioritizations, and are used for both in this study (Endeavour-GW and Pinta-GW for genome-wide prioritization, Endeavour-CS and Pinta-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, approaches (Köhler et al., 2008) and since they can be efficiently used with many training genes. The performance of Posmed shows a strong dependency on the set of keywords used as an input and we ran 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 is 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. Notice that Suspects went offline during our study after the 27th association and is not supported anymore (Euan Adie, personal communication), therefore, Suspects results are based on 27 associations over 42.

2.2 Validation data set

The validation data set is built by mining the scientific literature to identify the recently discovered disease-gene associations. This is achieved manually to avoid false positive associations. We select 6 journals that frequently 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 are confirmed or not. Novelty is assessed by using OMIM (McKusick, 1998), the Genetic Association Database (Becker et al., 2004), GoPubmed (Doms and Schroeder, 2005), and GeneCards (Safran et al., 2010). More precisely, we assess novelty at the gene level, and therefore novel mutations within already known genes are not considered. This process was kept active for 6 months (May 15 - November 15, 2010) and led to a collection of 42 associations (see Table 1

and Supplementary Table S2). For each association, the tools are run as soon as the association is identified following the defined workflow (see below). 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, Pinta 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 are collected from the corresponding OMIM pages, GAD pages and from recently published reviews when possible. 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 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.

The resulting 42 novel disease gene associations do not represent a homogeneous set. Therefore, we have divided them into confirmed (for monogenic diseases, the mutation is found in at least 2 unrelated patients; for multifactorial diseases, a GWAS is replicated in a separate cohort), intermediate (a single study, but additional functional evidence is provided), and unconfirmed (a single study) associations.

2.3 Performance measures

For each tool, we then assess its ability to identify the novel disease genes as promising genes using 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 of these rank ratios to give a more comprehensive view of the tool performance. Another 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. To compare the tools even further, we computed the true positive rates when setting the threshold for validation at the top 5% (TPR in top 5% of candidates), 10% (TPR in top 10%) and 30% (TPR in top 30%). This is motivated by the fact that in a real situation, the number of candidate genes to assay often needs to be limited because of financial and time constraints. We have selected three thresholds that represent reasonable biological hypotheses, as we previously illustrated in a genetic screen (Aerts et al., 2009). The corresponding TPR measures are used to estimate how efficient the tools are if only the top 5%, 10%, or 30% candidate genes would be assayed. Notice that these values correspond to the shape of the lower end of the ROC curve (the sharper the curve, the higher the TPR). There are cases for which some tools are not able to identify the novel disease gene at all, we therefore include a response rate. It is defined as the percentage of associations for which each tool does return a prioritization result for the novel disease gene (in some cases a tool will not return any result, for example because it could not correctly map the gene identifier or some candidates are otherwise filtered out). For example, if one of the 42 disease genes could not be ranked (i.e., gene is missing), the response rate drops down to ~98% (41/42).

Lastly, we also derive a heat map to detect any correlation between tools by computing the pairwise cosine similarity of the rankings presented in Tables 2 (see Supplementary Figure S1).

Table 1. The validation data set consisting of 42 recently discovered disease gene associations.

Gene	Disease/phenotype	Reference(s)
HCCS	Congenital Diaphragmatic Hernia	Qidwai et al. (2010)
BRCA2	Bipolar Disorder	Tesli et al. (2010)
TNFRSF19	Nasopharyngeal carcinoma	Bei et al. (2010)
MECOM	Nasopharyngeal carcinoma	Bei et al. (2010)
ATF7IP	Testicular germ cell tumor	Turnbull et al. (2010)
DMRT1	Testicular germ cell tumor	Turnbull et al. (2010)
FUT2	Crohn's disease	McGovern et al. (2010)
CSF1R	Asthma	Shin et al. (2010)
GLI3	Metopic craniosynostosis	McDonald-McGinn et al. (2010)
STOM	Nonsyndromic cleft lip/palate	Letra et al. (2010)
UTRN	Arthrogryposis	Tabet et al. (2010)
GABRR1	Bipolar schizoaffective disorder	Green et al. (2010)
UBE2L3	Crohn's disease	Fransen et al. (2010)
BCL3	Crohn's disease	Fransen et al. (2010)
EZH2	Myelodysplastic syndromes	Nikoloski et al. (2010)
TRAF6	Parkinson's disease	Zucchelli et al. (2010)
IL10	Behcet's disease	Remmers et al. (2010); Mizuki et al. (2010)
DAB2IP	Abdominal aortic aneurysm	Gretarsdottir et al. (2010)
SPIB	Primary biliary cirrhosis	Liu et al. (2010)
MMEL1	Primary biliary cirrhosis	Hirschfield et al. (2010)
TBX2	Complex heart defect	Radio et al. (2010)
RUNX2	Single-suture craniosynostosis	Mefford et al. (2010)
CRHR1	Multiple sclerosis	Briggs et al. (2010)
IFNG	Leprosy	Cardoso et al. (2010)
SH2B1	Congenital Anomalies of the Kidney and Urinary Tract	Sampson et al. (2010)
DISP1	Congenital Diaphragmatic Hernia	Kantarci et al. (2010)
G6PC3	Dursun syndrome	Banka et al. (2010)
PQBP1	Periventricular heterotopia	Sheen et al. (2010)
CD320	Methylmalonic aciduria	Quadros et al. (2010)
CHST14	Ehlers-Danlos syndrome	Miyake et al. (2010)
PLCE1	Esophageal squamous cell carcinoma	Wang et al. (2010); Abnet et al. (2010)
C20orf54	Esophageal squamous cell carcinoma	Wang et al. (2010)
SDCCAG8	Retinal-renal ciliopathy	Otto et al. (2010)
TP63	Lung adenocarcinoma	Miki et al. (2010)
UBE2E2	Type 2 diabetes	Yamauchi et al. (2010)
LPP	Tetralogy of Fallot	Arrington et al. (2010)
RANBP1	Smooth pursuit eye movement abnormality	Cheong et al. (2011)
HTR7	Alcohol dependence	Zlojutro et al. (2010)
SOX17	Congenital anomalies of the kidney and the urinary tract	Gimelli et al. (2010)
ACAD9	Mitochondrial complex I deficiency	Haack et al. (2010)
TRAF3IP2	Psoriasis	Ellinghaus et al. (2010); Hüffmeier et al. (2010)
WDR62	Autosomal recessive primary microcephaly	Yu et al. (2010); Nicholas et al. (2010)

2.4 Integration of predictions

In order to get an estimate of the usefulness of a meta-predictor, the results of the different tools are combined using the Order Statistics as within Endeavour. Integration happens separately for the genome-wide tools and for the candidate set based tools, and tools that return only few rankings (Suspects and Posmed) were not included. For each experiment, the gene identifiers of the different tools are mapped using Biomart. In order to avoid getting artificially favorable rankings, the size of the merged ranking is set to the maximum size of the underlying rankings.

3 RESULTS

The overall ranking results of all gene prioritization tools are summarized in Table 2, the complete results are presented in Supplementary Tables S9 and S10. These results have also been added to the Gene Prioritization Portal (http://www.esat.kuleuven.be/gpp).

Table 2. Results for the genome-wide and candidate set based prioritization tools. (*) Values computed only on the first 27 associations.

	Median	Response rate	TPR in top 5%	TPR in top 10%	TPR in top 30%
Genome-wide prioritization tools					
Candid	18.10	100%	21.4%	33.3%	64.3%
Endeavour-GW	15.49	100%	28.6%	38.1%	71.4%
Pinta-GW	19.03	100%	26.2%	31.0%	71.4%
Integration	12.45	100%	19.1%	38.1%	78.6%
Candidate set based prioritization tools					
Suspects	12.77*	88.9%*	33.3%*	33.3%*	63.0%*
ToppGene	16.80	97.6%	35.7%	42.9%	52.4%
GeneWanderer-RW	22.10	95.2%	16.7%	26.2%	61.9%
GeneWanderer-DK	22.97	88.1%	11.9%	21.4%	52.4%
Posmed-DN	45.45	50.0%	4.7%	11.9%	23.8%
Posmed-KS	31.44	47.6%	4.7%	7.1%	23.8%
GeneDistiller	11.11	97.6%	26.2%	47.6%	78.6%
Endeavour-CS	11.16	100%	26.2%	42.9%	90.5%
Pinta-CS	18.87	100%	28.6%	31.0%	71.4%
Integration	6.99	100%	40.5%	57.1%	83.3%

3.1 Performance measures

When considering the median of the rank ratios, GeneDistiller, Endeavour-CS, and Suspects are the tools that perform the best on this benchmark (respectively 11.11, 11.16, and 12.77). They are followed by Endeavour-GW (15.49), ToppGene (16.8), Candid (18.1), Pinta-CS (18.87), Pinta-GW (19.03), GeneWanderer-RW (22.11), GeneWanderer-DK (22.97), Posmed-KS (31.44), and Posmed-DN (45.45). The boxplots presented in Figure 1 illustrate that both, GeneDistiller and Endeavour-CS perform better than the other candidate set based prioritization tools (Figure 1-right). Among the genome-wide tools, Endeavour-GW performs slightly better than Pinta-GW and Candid (Figure 1-left).

When considering the response rate, Endeavour (both modes), Candid, and Pinta (both modes) performed the best study with 100% closely followed by ToppGene, GeneDistiller, and GeneWanderer-RW with more than 95% (meaning that only one or two associations are missing). At the other hand of the spectrum, Posmed-KS and Posmed-DN only work for about half of the experiments in our benchmark (respectively 47.6% and 50%).

When we compare the tools based on the global AUC (see Figure 2), we observe that GeneDistiller appears as the best performing tool overall with an AUC of 86%. It is followed by Endeavour-CS (82%), Endeavour-GW (79%), Pinta-GW (77%), Suspects (76%), Pinta-CS (75%), Candid (73%), GeneWanderer-RW (71%), GeneWanderer- DK (67%), ToppGene (66%), Posmed-KS (58%), and Posmed-DN (56%). However, the ROC curves are in general intertwined meaning that none of the approaches is clearly performing better than the other. However, we postulate that, in our case, the most important section of the ROC curve is the beginning and therefore use three other measures, the true positive rates at 5%, at 10%, and at 30%. These measures indicate how efficient the tools would be if only the top candidate genes would be assayed.

Considering the TPR in top 10% and 30%, we can observe a similar trend. Indeed, at 10%, GeneDistiller is first with a rate of 47.6% (20 associations found over 42), followed by both ToppGene and Endeavour-CS with 42.9% (18 associations). However, at 30%, the best tool is Endeavour-CS (90.5% - 38 associations), followed by GeneDistiller (78.6% - 33 associations). The other tools show smaller TPR at both levels: Pinta-CS (31%, 71.4%), Suspects (33.3%, 63%), GeneWanderer-RW (26.2%, 61.9%), GeneWanderer-DK (21.4%, 52.4%), Posmed-KS (7.1%, 23.8%), and Posmed-DN (11.9%, 23.8%). Among the genome-wide prioritization tools, Endeavour-GW shows highest TPR in top 10% and 30% (38.1%, 71.4%), followed by Candid (33.3%, 64.3%) and Pinta-GW (31%, 71.4%).

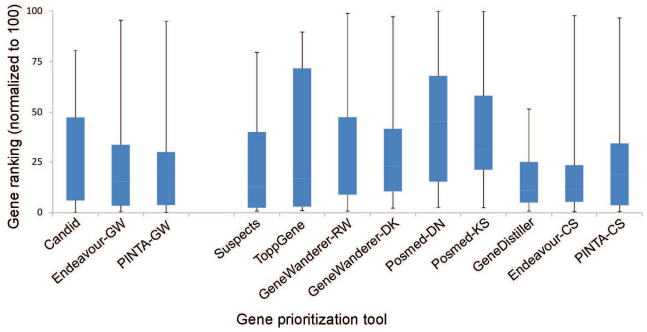


Figure 1: Boxplots of the 42 novel disease genes from the validation data set illustrated for the genome-wide (left) and candidate gene set based (right) prioritization tools.

3.2 Correlations

Supplementary Figure S1 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, Pinta), the two modes are highly correlated (> 0.89). There is also a significant correlation between Candid and GeneWanderer-DK (0.82). The other values are within 0.4 and 0.7, indicating that all tools are moderately correlated.

3.3 Integration of predictions

Our meta-analysis reveals that the best results are obtained when predictions are combined over the different tools (see Table 2 and Supplementary Table S11). For the genome-wide tools, all performance measures are improved by the integrative method (e.g.,

median of 12.45 for the meta-predictor versus 15.49 for Endeavour-GW). Similar results are obtained for the candidate set based tools (e.g., median of 6.99 for the meta-predictor versus 11.11 for GeneDistiller), although the TPR in the top 30% of the integrative method is still lower than for Endeavour-CS (83.3% versus 90.5%).

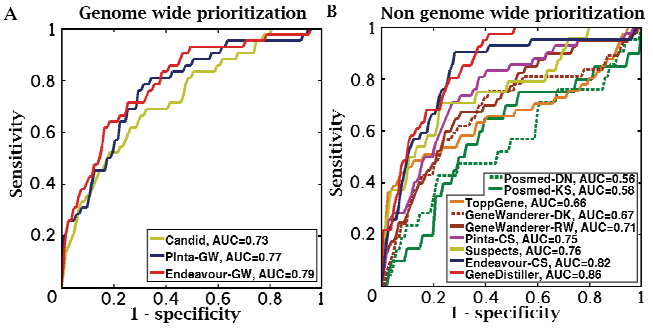


Figure 2: ROC curves of the genome-wide (A) and candidate gene set based (B) prioritization tools.

4 DISCUSSION

We aim at assessing the usefulness of eight gene prioritization tools that are freely available via web applications. We have built a validation based on 42 recently discovered disease-gene associations from literature containing novel genes for both monogenic conditions and complex disorders. We have selected novel disease-gene 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 so that all databases used for gene prioritization are still not contaminated by the knowledge of the novel disease-gene association. This also means that we had to exclude tools that query MEDLINE online since their results would be biased.

We want to compare the performance of the tools even if the inputs are different (genes vs. keywords, genome-wide vs. candidate set). Among the eight gene prioritization tools that we have analyzed in this study, only Endeavour, Candid, and Pinta have been used for genome-wide prioritization. The input data for Endeavour and Pinta are training genes, whereas Candid requires keywords. The gene prioritization tools that we have used to prioritize candidate genes within a region of interest are Suspects, ToppGene, GeneWanderer, Posmed, GeneDistiller, and again Endeavour and Pinta. Suspects and Posmed are trained with keywords, the other tools require 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 set of appropriate keywords to derive fair and meaningful comparisons. However, different, and possibly better, results might be obtained 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, GeneDistiller and Endeavour-CS consistently appear as the best tools when looking at all performance measures. It is interesting to notice that the best results are in general obtained with tools that use many data types in conjunction (up to eight for Endeavour, as compared to the three data sources used by Posmed), but there is no perfect correlation. This is in agreement with the conclusion of the recent review by Tiffin et al. (2009), who indicate that successful computational applications will be facilitated by improved data integration.

All tools except Posmed have a high response rate ranging from 88% to 100%, meaning that at least 37 of the 42 novel disease genes are prioritized (or 24 of 27 for Suspects). However, the response rates for Posmed-KS and Posmed-DN are respectively 47.6% and 50%, which can be explained by the fact that Posmed also acts as a filter on the candidate genes to obtain a reduced list of genes in the end. There are therefore cases for which the novel disease gene has been removed by the filter. This is different from the other tools for which missing genes basically correspond to genes that are not recognized by the tool (it happens most of the time with poorly characterized genes, such as C20orf54). Another special case is Suspects that went offline during the validation and therefore could only be validated with the first 27 associations. We therefore calculated the response rate only on the first 27 associations.

Two types of tools can be distinguished, the ones that are trained with already known genes and the ones that are trained with descriptive keywords. It appears that gene-based tools seem to work better than keyword-based tools (the average of medians is 17.2 for genes based tools and 27 for keyword based tools - similar results are obtained with the other measures, see Supplementary Table S8). This could be because we use in general more genes than keywords for training (18.8 genes on average for 6 keywords). This also indicates that more keywords might be needed to model a disease, a small text (such as an OMIM entry) might even be necessary (van Driel et al., 2006).

There is in general an agreement between the five performance measures we use throughout our study. One notable exception exists for ToppGene, whose AUC is 66%, and corresponds to rank 10th (out of the 12 prioritization tools). In contrast its associated TPR in top 10% is 42.9%, which corresponds to rank 2nd. This apparent contradiction can be explained by observing Figure 2, in which the ROC curve exhibits a non convex shape. This is because ToppGene either ranks the novel disease gene on top or at the bottom (i.e., the disease genes are rarely ranked in the middle). And therefore the TPR in top 10% will be high because it only takes into account the top of the list, while the AUC will be lower because it basically behaves like an average over all cases. Another important point is that our observations are in line with the 'no free lunch' theorem. Indeed each tool can perform better than all the others for some cases, or, in other words, none of the tools outperforms another on the complete data set (if we do not consider the special case of Posmed that also acts as a filter).

Posmed-KS has been trained with the complete keyword set, whereas Posmed-DN has been trained only with the disease name. The median rank ratio is 31.44 when the complete keyword set is used and drops to 45.45 when only the disease name is inputted. If we only compare the results over the 19 associations for which both tools are able to prioritize the novel disease gene, the difference becomes even larger (29.6 and 50 respectively for Posmed-KS and Posmed-DN). Altogether, these results indicate that Posmed does not rely on the use of the single disease name and that the extra keywords are indeed important. It can be observed that the performance measures for Posmed are worse than for the other tools in our benchmark study. However, when looking at the individual ranks, it can be observed that Posmed returns far fewer genes than the other tools because it also acts as a filter. As a result, the rank ratios are on general larger and the performance

measures are therefore worse. As such, it becomes difficult to fairly compare Posmed to the other tools because our measures of performance naturally penalize the fact that Posmed returns prioritizations for a limited set of candidates. Changing our performance measures to counterbalance this effect would then give an unfair advantage to Posmed because it returns prioritizations only for the "safer bets".

GeneWanderer has also been run twice with different network algorithms: random walk (RW) and diffusion kernel (DK). The respective performance are very similar although the random walk approach is performing a little bit better than the diffusion kernel albeit non significant (22.11 to 22.97 for median rank ratio – similar differences are observed with the other measures). The heat map indicates a strong correlation (>0.9, see Supplementary 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 algorithms are similar.

Endeavour and Pinta are used to prioritize both the whole genome (Endeavour-GW and Pinta-GW) and the defined chromosomal region (Endeavour-CS and Pinta-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 (11.16) than for Endeavour-GW (15.49) in our benchmark. The difference is smaller but remains when considering the AUC, and the TPR in top 10% and 30%.

The same training genes are 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 Supplementary Figure S1). At contrary, the results for both Pinta modes are very similar (correlation of 0.99) and seem to indicate that the size of the candidate set does not influence this algorithm.

In this study, we consider the tools as off the shelf solutions, and use them as recommended by the developers without fine tuning of the parameters. However, 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, when compared to data that is two years old. In our setup, we have no control over the genomic data used and cannot identify if variation in performance among tools can be explained by this.

In addition, the quality of both the data sources and the integration methodologies are also influencing the outcome of the prioritization process. However, we aim at estimating the usefulness of some prioritization tools for geneticists. And therefore an in-depth comparison of the implementation of the tools is beyond the scope of this study.

It is important to notice that the 42 novel disease gene associations do not represent a very homogeneous set. Indeed, the median of the rank ratios over the tools show that some associations seem to be easier to predict than others. This also explains why all tools are moderately correlated on the heat map (> 0.4). A plausible explanation is the disparity in the available data between the novel disease genes. Since only little data can be gathered for poorly characterized genes, such as C20orf54, they are more difficult to prioritize. However, we also hypothesize that the nature of the underlying genetic disorder, as well as the quality of the reported association might influence the ability of the tools to predict correctly that association. We have therefore divided the associations between confirmed, intermediate, and unconfirmed. Among the 42 associations, 23 are confirmed, 8 are intermediate, and 11 are unconfirmed (see Supplementary Table S2). We hypothesize that this might influence our validation since some unconfirmed associations might in fact be spurious. We observe that Suspects and ToppGene perform better for the 23 confirmed associations than for the 19 unconfirmed ones (see Supplementary Tables S4 and S5). This trend is however not always shared as the situation is opposite for GeneDistiller and GeneWanderer. Although informative, these comparisons are not significant due to the small number of associations.

In our validation data set, there are 17 monogenic diseases and 25 multifactorial disorders (see Supplementary Tables S6 and S7). It has been shown that it is more difficult to make predictions for multifactorial diseases than for monogenic diseases (Linghu et al., 2009). Our results however seem to indicate that not all tools are influenced by the intrinsic complexity of multifactorial diseases. For instance, Endeavour and ToppGene seem to perform better for monogenic conditions while GeneWanderer and Suspects perform better for complex disorders. However, the size of our validation data set does not allow for a complete statistical analysis. Larger validation data sets and real predictive studies will be pursued to complement our preliminary study.

Several studies have shown that combining predictions of several tools lead to even better predictions (Tiffin et al., 2006; Elbers et al., 2007). However, no performance criteria were used to select the tools to be combined. With this comparison of tools, we ease the selection of the most efficient tools, whose combination may lead to more accurate predictions. In addition, we report that the meta-predictors that integrate the predictions made by several tools perform better than the best individual tools as already reported (Thornblad et al., 2007).

Our results indicate that cross-validation based benchmarks tend to overestimate the real predictive performance. Indeed, all the tools for which such a benchmark exists have lower AUC than anticipated using our dataset (see Supplementary Table S12). We therefore believe that developers should take extra care when benchmarking their tools as to avoid these pitfalls. Also, some hard constraints have made this study small enough not to reach significance (e.g., only few tools have a programmatically queryable interface).

As already discussed in (Moreau et al., 2012), this field needs to consolidate through improved benchmarking efforts due to the lack of a ground truth for evaluating the performance of prioritization methods. Therefore we see a need for a large-scale community effort to compare multiple tools across common prospective benchmarks. We hope our work represents the first step towards a collaborative effort to tackle this problem at a larger scale.

ACKNOWLEDGEMENTS

We thank Peter Konings for his help regarding the statistics.

Funding:

Research Council KUL [CIF/07/02 DE CAUSMAE / DEFIS - SOCK, ProMeta, GOA Ambiorics, GOA MaNet, GOA 2006/12, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys, START 1, several PhD/postdoc and 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

Abnet, C. C. et al., (2010). A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat Genet, 42(9), 764–767.

Adie, E. A. et al., (2006). SUSPECTS: enabling fast and effective prioritization ofpositional candidates. Bioinformatics, 22(6), 773 –774.

Aerts, S. et al., (2006). Gene prioritization through genomic data fusion. Nat Biotech, 24(5), 537–544.

Aerts, S. et al., (2009). Integrating computational biology and forward genetics in drosophila. PLoS Genet, 5(1), e1000351.

Arrington, C. B. et al., (2010). 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, 152A(11), 2919–2923. PMID: 20949626.

Banka, S. et al., (2010). Mutations in the G6PC3 gene cause dursun syndrome. American Journal of Medical Genetics. Part A, 152A(10), 2609–2611. PMID: 20799326.

Becker, K. G. et al., (2004). The genetic association database. Nat Genet, 36(5), 431–432.

Bei, J. et al., (2010). A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat Genet, 42(7), 599–603.

Briggs, F. B. S. et al., (2010). Evidence for CRHR1 in multiple sclerosis using supervised machine learning and meta-analysis in 12,566 individuals. Human Molecular Genetics, 19(21), 4286–4295. PMID: 20699326.

Calvo, S., et al. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. Nat Genet, 38(5):576-82.

Cardoso, C. C. et al., (2010). IFNG +874 TntextgreaterA single nucleotide polymorphism is associated with leprosy among brazilians. Human Genetics, 128(5), 481–490. PMID: 20714752.

Chen, J. et al., (2007). Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics, 8(1), 392.

Cheong, H. S. et al., (2011). 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, 156B(1), 67–71. PMID: 21184585.

Doms, A. and Schroeder, M. (2005). GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Research, 33(Web Server issue), W783–786. PMID: 15980585.

Doncheva, N. T., et al., (2012). Recent approaches to the prioritization of candidate disease genes. WIREs Syst Biol Med. doi: 10.1002/wsbm.1177.

Elbers, C. C. et al., (2007). A strategy to search for common obesity and type 2 diabetes genes. Trends in Endocrinology and Metabolism: TEM, 18(1), 19–26. PMID: 17126559.

Ellinghaus, E. et al., (2010). Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. Nat Genet, 42(11), 991–995.

Erlich, Y. et al., (2011). Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res. 21: 658-664.

Fransen, K. et al., (2010). 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, 19(17), 3482–3488. PMID: 20601676.

Gimelli, S. et al., (2010). Mutations in SOX17 are associated with congenital anomalies of the kidney and the urinary tract. Human Mutation, 31(12), 1352–1359. PMID: 20960469.

Green, E. K. et al., (2010). 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, 153B(7), 1347–1349. PMID: 20583128.

Gretarsdottir, S. et al., (2010). Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. Nature Genetics, 42(8), 692–697. PMID: 20622881.

Haack, T. B. et al., (2010). Exome sequencing identifies ACAD9 mutations as a cause of complex i deficiency. Nat Genet, 42(12), 1131–1134.

Haider, S. et al., (2009). BioMart central portal–unified access to biological data. Nucleic Acids Research, 37(Web Server), W23–W27.

Hardy, J. and Singleton, A. (2009). Genomewide association studies and human disease. The New England Journal of Medicine, 360(17), 1759–1768. PMID: 19369657.

Hirschfield, G. M. et al., (2010). Variants at IRF5-TNPO3, 17q12-21 and MMEL1 are associated with primary biliary cirrhosis. Nat Genet, 42(8), 655–657.

Hüffmeier, U. et al., (2010). Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis. Nature Genetics, 42(11), 996–999. PMID: 20953186.

Hutz, J. E. et al., (2008). CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genetic Epidemiology, 32(8), 779–790. PMID: 18613097.

Kantarci, S. et al., (2010). Characterization of the chromosome 1q41q42.12 region, and the candidate gene DISP1, in patients with CDH. American Journal of Medical Genetics. Part A, 152A(10), 2493–2504. PMID: 20799323.

Köhler, S. et al., (2008). Walking the interactome for prioritization of candidate disease genes. American Journal of Human Genetics, 82(4), 949–958. PMID: 18371930.

Letra, A. et al., (2010). Follow-up association studies of chromosome region 9q and nonsyndromic cleft lip/palate. American Journal of Medical Genetics. Part A, 152A(7), 1701–1710. PMID: 20583170.

Linghu, B. et al., (2009). Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biology, 10(9), R91. PMID: 19728866.

Liu, X. et al., (2010). Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. Nature Genetics, 42(8), 658–660. PMID: 20639880.

Lupski, J. R. et al., (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. The New England Journal of Medicine, 362(13), 1181–1191. PMID: 20220177.

McDonald-McGinn, D. M. et al., (2010). Metopic craniosynostosis due to mutations in GLI3: a novel association. American Journal of Medical Genetics. Part A, 152A(7), 1654–1660. PMID: 20583172.

McGovern, D. P. B. et al., (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with crohn's disease. Human Molecular Genetics, 19(17), 3468–3476. PMID: 20570966.

McKusick, V. A. (1998). Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders. The Johns Hopkins University Press, 12th edition.

Mefford, H. C. et al., (2010). 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, 152A(9), 2203–2210. PMID: 20683987.

Miki, D. et al., (2010). Variation in TP63 is associated with lung adenocarcinoma susceptibility in japanese and korean populations. Nat Genet, 42(10), 893–896.

Miyake, N. et al., (2010). Loss-of-function mutations of CHST14 in a new type of Ehlers-Danlos syndrome. Human Mutation, 31(8), 966–974. PMID: 20533528.

Mizuki, N. et al., (2010). Genome-wide association studies identify IL23R-IL12RB2 and IL10 as behc¸et's disease susceptibility loci. Nature Genetics, 42(8), 703–706. PMID: 20622879.

Moreau Y., et al. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discover. Nat Rev Genet.13(8):523-36.

Nicholas, A. K. et al., (2010). WDR62 is associated with the spindle pole and is mutated in human microcephaly. Nat Genet, 42(11), 1010–1014.

Nikoloski, G. et al., (2010). Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. Nat Genet, 42(8), 665–667.

Nitsch, D. et al., (2010). Candidate gene prioritization by network analysis of differential expression using machine learning approaches. BMC Bioinformatics, 11(1), 460.

Oti, M. (2011). Web tools for the prioritization of candidate disease genes. Methods Mol Biol.760:189-206.

Otto, E. A. et al., (2010). Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. Nat Genet, 42(10), 840–850.

Piro, R. M. et al., (2012). Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS Journal, 279: 678–696.

Qidwai, K. et al., (2010). 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, 152A(6), 1588–1590. PMID: 20503342.

Quadros, E. V. et al., (2010). Positive newborn screen for methylmalonic aciduria identifies the first mutation in TCblR/CD320, the gene for cellular uptake of transcobalamin-bound vitamin b(12). Human Mutation, 31(8), 924–929. PMID:20524213.

Radio, F. C. et al., (2010). TBX2 gene duplication associated with complex heart defect and skeletal malformations. American Journal of Medical Genetics. Part A, 152A(8), 2061–2066. PMID: 20635360.

Rajab, A., et al. (2010). Fatal Cardiac Arrhythmia and Long-QT Syndrome in a New Form of Congenital Generalized Lipodystrophy with Muscle Rippling (CGL4) Due to *PTRF-CAVIN* Mutations. PLoS Genet 6(3): e1000874.

Remmers, E. F. et al., (2010). Genome-wide association study identifies variants in the MHC class i, IL10, and IL23R-IL12RB2 regions associated with behc¸et's disease. Nat Genet, 42(8), 698–702.

Safran, M. et al., (2010). GeneCards version 3: the human gene integrator. Database: The Journal of Biological Databases and Curation, 2010, baq020. PMID: 20689021.

Sampson, M. G. et al., (2010). 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, 152A(10), 2618–2622. PMID: 20799338.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. Nat Meth, 5(1), 16–18.

Seelow, D. et al., (2008). GeneDistiller—Distilling candidate genes from linkage intervals. PLoS ONE, 3(12), e3874.

Sheen, V. L. et al., (2010). Mutation in PQBP1 is associated with periventricular heterotopia. American Journal of Medical Genetics. Part A, 152A(11), 2888–2890. PMID: 20886605.

Shin, E. K. et al., (2010). Association between colony-stimulating factor 1 receptor gene polymorphisms and asthma risk. Human Genetics, 128(3), 293–302. PMID: 20574656.

Tabet, A. et al., (2010). Molecular characterization of a de novo 6q24.2q25.3 duplication interrupting UTRN in a patient with arthrogryposis. American Journal of Medical Genetics Part A, 152A(7), 1781–1788.

Teber, E. T. et al., (2009). Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies. BMC Bioinformatics, 10 Suppl 1, S69. PMID: 19208173.

Tesli, M. et al., (2010). 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, 153B(7), 1276–1282. PMID: 20872766.

Thienpont B. et al., (2010). Haploinsufficiency of TAB2 causes congenital heart defects in humans. Am. J. Hum. Genet. 86(6): 839-849.

Thornblad, T. A. et al., (2007). 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, 10(6), 861–870. PMID: 18179399.

Tiffin, N. et al., (2006). Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Research, 34(10), 3067–3081. PMID: 16757574

Tiffin, N. et al., (2009). Linking genes to diseases: it's all in the data. Genome Medicine, 1(8), 77. PMID: 19678910.

Tiffin, N. (2011). Conceptual thinking for in silico prioritization of candidate disease genes. Methods Mol Biol.760:175-87.

Tranchevent, L. et al., (2010). A guide to web tools to prioritize candidate genes. Briefings in Bioinformatics.

Turnbull, C. et al., (2010). Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. Nature Genetics, 42(7), 604–607.

van Driel, M. A. et al., (2006). A text-mining analysis of the human phenome. European Journal of Human Genetics: EJHG, 14(5), 535–542. PMID: 16493445.

Vliet-Ostaptchouk, J. V. et al., (2008). HHEX gene polymorphisms are associated with type 2 diabetes in the Dutch Breda cohort, Eur J Hum Genet. 16, 652–656.

Wang, L. et al., (2010). Genome-wide association study of esophageal squamous cell carcinoma in chinese subjects identifies susceptibility loci at PLCE1 and c20orf54. Nat Genet, 42(9), 759–763.

Yamauchi, T. et al., (2010). A genome-wide association study in the japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4AC2CD4B. Nat Genet, 42(10), 864–868.

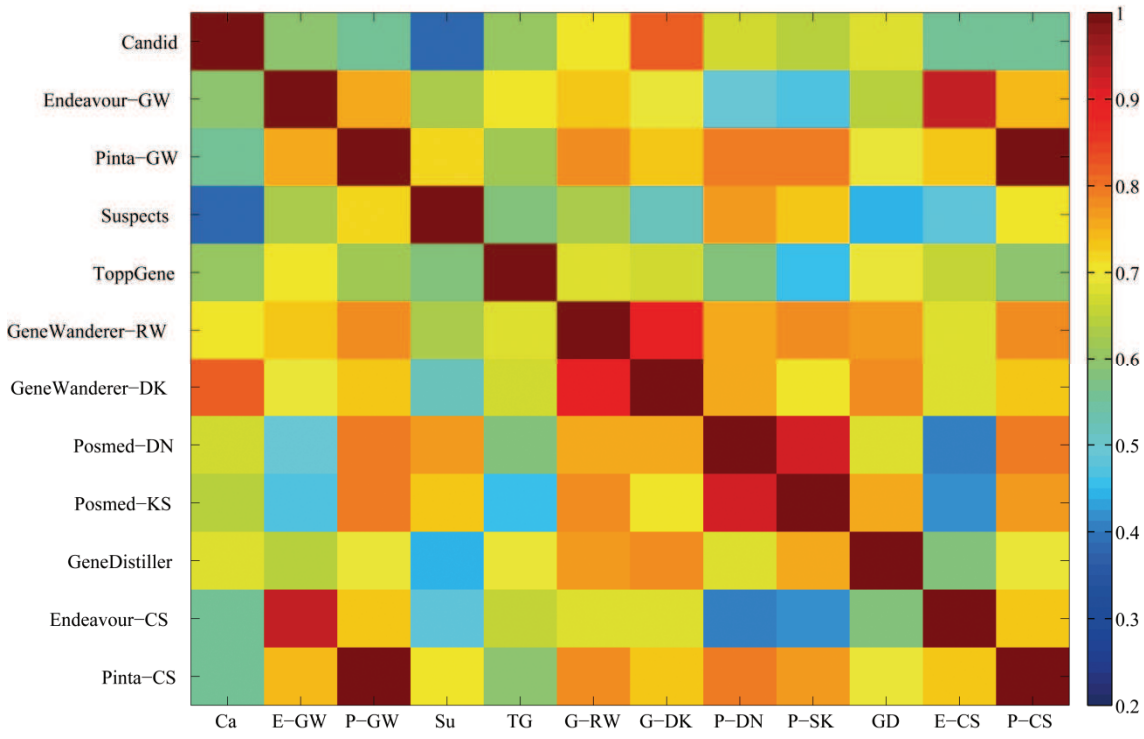
Yoshida, Y. et al., (2009). PosMed (Positional medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. Nucleic Acids Research, 37(Web Server issue), W147–152. PMID: 19468046.

Yu, T. W. et al., (2010). Mutations in WDR62, encoding a centrosome-associated protein, cause microcephaly with simplified gyri and abnormal cortical architecture. Nat Genet, 42(11), 1015–1020.

Zlojutro, M. et al., (2010). 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. PMID: 21046636.

Zucchelli, S. et al., (2010). 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, 19(19), 3759–3770. PMID: 20634198.

SUPPLEMENTARY FIGURE S1: HEAT MAP OF THE GENE PRIORITIZATION METHODS USED IN THIS STUDY.



SUPPLEMENTARY TABLE S1: OVERVIEW OF THE GENE PRIORITIZATION METHODS.

GP Tool	Publication	Website
Suspects	Adie et al. (2006)	http://www.genetics.med.ed.ac.uk/suspects/
ToppGene	Chen et al. (2007)	http://toppgene.cchmc.org/
GeneDistiller	Seelow et al. (2008)	http://www.genedistiller.org/
GeneWanderer	Köhler et al. (2008)	http://compbio.charite.de/genewanderer/GeneWanderer
Posmed	Yoshida et al. (2009)	http://omicspace.riken.jp/
Candid	Hutz et al. (2008)	https://dsgweb.wustl.edu/hutz/candid.html
Endeavour	Aerts et al. (2006)	http://www.esat.kuleuven.be/endeavour
Pinta	Nitsch et al. (2011)	http://www.esat.kuleuven.be/pinta

SUPPLEMENTARY TABLE S2: LIST OF THE 42 NOVEL DISEASE GENE ASSOCIATIONS.

Novel Gene	Disease	Inheritance	Type of study	Finding	Status	Confirmed associations: proof
ACAD9	Mitochondrial complex I deficiency	monogenic	patient report	mutation	confirmed	4 unrelated patients.
CHST14	Ehlers-Danlos syndrome	monogenic	patient report	mutation	confirmed	6 unrelated patients.
EZH2	Myelodysplastic syndromes	monogenic	mutation screen	deletion, mutation	confirmed	22 unrelated patients.
G6PC3	Dursun syndrome	monogenic	case report	mutation	confirmed	2 unrelated patients.
GLI3	Metopic craniosynostosis	monogenic	patient report	mutation	confirmed	2 unrelated patients.
HCCS	Congenital Diaphragmatic Hernia	monogenic	patient report	deletion	confirmed	4 unrelated patients. CDH can be a feature of MLS.
SDCCAG8	Retinal-renal ciliopathy	monogenic	patient report	mutation	confirmed	8 unrelated patients.
SOX17	Congenital anomalies of the kidney and the urinary tract	monogenic	patient report	mutation	confirmed	8 unrelated patients.
WDR62	autosomal recessive primary microcephaly	monogenic	GW linkage analysis	SNP	confirmed	2 unrelated families.
CRHR1	Multiple sclerosis	multifactorial	GWAS	SNP	confirmed	Replication in independent cohort and meta-analysis.
DAB2IP	Abdominal aortic aneurysm	multifactorial	GWAS	SNP	confirmed	Replication in independent cohort and meta-analysis.
DMRT1	Testicular germ cell tumor	multifactorial	GWAS	SNP	confirmed	Kanetsky et al. (2011) Hum Mol Genet. 20(15):3109.
FUT2	Crohn's disease	multifactorial	GWAS	SNP	confirmed	Replication in independent cohort and GWAS meta-analysis.
GABRR1	Bipolar schizoaffective disorder	multifactorial	GWAS	SNP	confirmed	Wang et al. (2010) Schizophr Res. 124(1-3):192.
IFNG	Leprosy	multifactorial	GWAS	SNP	confirmed	Replication in independent cohort.
IL10	Behçet's disease	multifactorial	GWAS	SNP	confirmed	Replication in independent cohort and meta-analysis.
MMEL1	Primary biliary cirrhosis	multifactorial	GWAS	SNP	confirmed	Replication in independent cohort and meta-analysis.
RANBP1	smooth pursuit eye movement abnormality	multifactorial	SNP geno-typing	SNP	confirmed	Cheong et al. (2011) Am J Med Genet B Neuropsychiatr Genet. 156B(1):67.
PLCE1	Esophageal squamous cell carcinoma	multifactorial	GWAS	SNP	confirmed	Two unrelated studies.
SPIB	Primary biliary cirrhosis	multifactorial	GWAS	SNP	confirmed	Tanaka et al. (2011) Tissue Antigens. 78(1):65.
TP63	Lung adenocarcinoma	multifactorial	GWAS	SNP	confirmed	Hu et al. (2011) Nat Genet.43(8):792.
TRAF3IP2	Psoriasis	multifactorial	GWAS	SNP	confirmed	Two unrelated studies.

UBE2E2	Type 2 diabetes	multifactorial	GWAS	SNP	confirmed	Replication in two independent cohorts.
CD320	Methylmalonic aciduria	monogenic	Patient report	deletion	intermediate	
DISP1	Congenital Diaphragmatic Hernia	monogenic	GWAS	deletion, micro-deletion, SNP, mutation	intermediate	
PQBP1	Periventricular heterotopia	monogenic	Patient report	deletion	intermediate	
UTRN	Arthrogryposis	monogenic	Patient report	mutation	intermediate	
C20orf54	Esophageal squamous cell carcinoma	multifactorial	GWAS	SNP	intermediate	
HTR7	alcohol dependence	multifactorial	GWAS	SNP	intermediate	
STOM	Nonsyndromic cleft lip/palate	multifactorial	GWAS	SNP	intermediate	
UBE2L3	Crohn's disease	multifactorial	GWAS	SNP	intermediate	
LPP	Tetralogy of Fallot	monogenic	Patient report	deletion	weak	
RUNX2	Single-suture craniosynostosis	monogenic	mutation screen	duplication	weak	
SH2B1	Congenital Anomalies of the Kidney and Urinary Tract	monogenic	case report	Micro-deletion	weak	
TBX2	Complex heart defect	monogenic	case report	duplication + model organism data	weak	
ATF7IP	Testicular germ cell tumor	multifactorial	GWAS	SNP	weak	
BCL3	Crohn's disease	multifactorial	GWAS	SNP	weak	
BRCA2	Bipolar schizoaffective disorder	multifactorial	GWAS	SNP	weak	
CSF1R	Asthma	multifactorial	GWAS	SNP	weak	
MECOM	Nasopharyngeal carcinoma	multifactorial	GWAS	SNP	weak	
TNFRSF19	Nasopharyngeal carcinoma	multifactorial	GWAS	SNP	weak	
TRAF6	Parkinson's disease	multifactorial	Hypo-thesis	expression and interaction	weak	

SUPPLEMENTARY TABLE S3: INPUT DATA FOR THE 42 EXPERIMENTS.

Gene	Keywords	Candidate region	Training genes	Tissues
HCCS	Congenital diaphragmatic hernia Diaphragm Pulmonary hypoplasia Pulmonary hypertension	Xp22.31-Xp22.2 (6000001-17100000) → 11.1 Mb	GATA4 RBP1 RBP2 NR2F2 WT1 SLIT3 STRA6 PDGFRA	Lung Fetal lung
BRCA2	Bipolar disorder Major affective disorder Manic-depressive Mania Depression	13q12.3-q13.3 (28900001-40100000) → 11.2 Mb	PALB2 SLC6A3 HTR4 ABCA13 DRD4 BDNF CUX2 SLC6A4 BCR COMT XBP1 TRPM2	Temporal lobe Globus pallidus Cerebellum peduncles Cerebellum Caudate nucleus Whole brain Parietal lobe Medulla oblongata Amygdala Prefrontal cortex Occipital lobe Hypothalamus Thalamus Subthalamic nucleus Cingulate cortex Fetal brain Adrenal cortex
TNFRSF19	Nasopharyngeal carcinoma Cancer Neoplasm Epithelium Nasopharynx Epstein-Barr Pharynx	13q12 (18397582 - 31095913) → 12.7 Mb	COX7B2 LOC344967 PLUNC TP53 HLA-A GABBR1 HLA-F CCND1 CYP2A6 XRCC1 CYP2F1 UBAP1 RASSF1A HHATL CDH13 CTLA-4 CYP2E1 ERCC1 FAS GSTM1 HP OGG1 HSPA1B IFNA17 IL10 IL12A IL16 IL18 IL1B IL8 MDM2 MICA MMP1 MMP2 N4BP2	Colorectal adenocarcinoma Bronchial epithelial cells

			NAT2 NFKB1 TAP1 TGFB1 TLR10 TLR3 TLR4 TNF VEGFA XPC HLA-B HLA-C HLA-E HLA-DPB1 HLA-DQA1 HLA-DQB1 HLA-DRB1 ITGA9	
MECOM	See TNFRSF19	3q26.1 - q26.2 (162152104- 172748973) → ~10Mb	See TNFRSF19	See TNFRSF19
ATF7IP	Testicular germ cell tumor Testicular Germ cell Tumor Seminoma Testis Carcinoma Teratoma Testicular microlithiasis Subfertility	12p13 (0 - 14682211) → 16.6 Mb	KITLG SF MGF SCF FPH2 SPRY4 BAK1 BAK CDN1 STK11 PJS LKB1 KIT PBT SCFR FGFR3 ACH CEK2 JTK4 TGFB1 CED LAP DPD1 TGFB LTA LT TNFB TNFSF1 TNF	Testis Testis - Leydig cell Testis - germ cell Testis - interstitial cell Testis - seminiferous tubule
DMRT1	See ATF7IP	9p23-p24 (0 - 14287822) → 14.4 Mb	See ATF7IP	See ATF7IP
FUT2	Crohn's disease Inflammatory bowel disease Gastrointestinal inflammation Commensal flora	19q13.33 - 19q13.43 (48000001 - 59128983) → 11.1 Mb	NOD2 IL23R ATG16L1 MST1 PTGER4 IRGM TNFSF15 ZNF365 NKX2-3 PTPN2 PTPN22 ITLN1	Colorectal adenocarcinoma Liver Kidney Pancreas Salivary gland Tongue Colon Small intestine

			IL12B CDKAL1 CCR6 JAK2 C11orf30 LRRK2 MUC19 ORMDL3 STAT3 ICOSLG	
CSF1R	Asthma Inflammatory disease Airways Bronchospasm Airflow obstruction Lungs Bronchi	5q31.3 - 5q32 (139500001-149800000) → ~10Mb	GSTM1 FLG IL10 CTLA4 IL13 IL4 CD14 SPINK5 ADRB2 HAVCR1 LTC4S LTA TNF HLA-DRB1 HLA-DQB1 HLA-DPB1 GPRA NAT2 CC16 GSTP1 IL18 STAT6 NOS1 CMA1 IL4R CCL11 CCL5 ACE TBXA2R TGFB1 ADAM33 GSTT1	Lung Fetal lung Bronchial epithelial cells Trachea Tongue Olfactory bulb
GLI3	Metopic craniosynostosis Craniosynostosis Trigonocephaly Polysyndactyly Brain Limb	7p14.1 - p13 (35000001-45400000) → 13Mb	FGFR1 FGFR2 FGFR3 TWIST1	Temporal lobe Globus pallidus Cerebellum peduncles Cerebellum Caudate nucleus Whole brain pPrietal lobe Medulla oblongata Amygdala Prefrontal cortex Occipital lobe Hypothalamus Thalamus Subthalamic nucleus Cingulate cortex Fetal brain Adrenal cortex
STOM	Cleft lip Cleft palate Craniofacial defect	9q33.1 - q33.3 (117700001-130300000) → 12 Mb	PVRL1 MSX1 IRF6 TP63 BMP4 SUMO1 MTR	None

UTRN	Arthrogryposis Hypertelorism Downslanting palpebral fissures Tented upper lip Short neck Severe mental retardation Growth retardation Joint contractures	6q24.1-q24.3 (139000001-149000000) → ca. 10Mb	TNNT3 TNNI2 MYH3 TPM2	Fetal liver Bone marrow Whole brain Fetal brain Fetal thyroid Fetal lung
GABRR1	Bipolar disorder Major affective disorder Mania Schizophrenia Bipolar mood episodes	6q15-q16.1 (88000001-99500000) → 11.5 Mb	PALB2 SLC6A3 HTR4 ABCA13 DRD4 BDNF CUX2 SLC6A4 BCR COMT XBP1 TRPM2	Temporal lobe Globus pallidus Cerebellum peduncles Cerebellum Caudate nucleus Whole brain Parietal lobe Medulla oblongata Amygdala Prefrontal cortex Occipital lobe Hypothalamus Thalamus Subthalamic nucleus Cingulate cortex Fetal brain Adrenal cortex
UBE2L3	Crohn's disease Inflammatory bowel disease Gastrointestinal inflammation Commensal flora	22q11.21 - 22q12.1 (17900001-29600000) → 11.7 Mb	NOD2 IL23R ATG16L1 MST1 PTGER4 IRGM TNFSF15 ZNF365 NKX2-3 PTPN2 PTPN22 ITLN1 IL12B CDKAL1 CCR6 JAK2 C11orf30 LRRK2 MUC19 ORMDL3 STAT3 ICOSLG FUT2	Whole blood Pancreas Pancreatic islets
BCL3	See UBE2L3	19q13.2 - 19q13.33 (3870000-5140000) → 12.7 Mb	See UBE2L3	See UBE2L3
EZH2	Myelodysplastic Leukemia Hematological Dysplasia Myeloid Bone marrow Blood	7q36.1-36.3 (147479482-158821424) → 11Mb	RPS14 TET2 HAX1 CSF3R ETV6 CEBPA NPM1 HPSE NQO1 GSTT1 CYP3A4 GSTM1	CD33+ myeloid cells CD34+ cells Lymphoma - Burkitt's (Raji) Leukemia - promyelocytic (hl60) Lymphoma - Burkitt's (Daudi) Leukemia - chronic Myelogenous (k562) Bone marrow

			CSF1R ASXL1 FLT3 HIPK2 KRAS2 JAK2	
TRAF6	Parkinson's disease Neurogenerative disorder Dopaminergic neurons Substantia nigra Lewy bodies	11p12 - 11p11.2 (36400001-48800000) \ → 12.4 Mb	SNCA Parkin PARK3 PARK4 UCHL1 PINK1 DJ-1 LRRK2 ATP13A2 GIGFY2 PARK10 PARK12 HTRA2	Temporal lobe Globus pallidus Cerebellum peduncles Cerebellum Caudate nucleus Whole brain Parietal lobe Medulla oblongata Amygdala Prefrontal cortex Occipital lobe Hypothalamus Thalamus Subthalamic nucleus Cingulate cortex Fetal brain
IL10	Behcet Inflammatory Immune Ulcer Skin lesion Eye Genital Oral	1q32.1 - 1q32.3 (197338641-211937672) → 14Mb	MICA HLA-A HLA-B MEFV SOD2 CYP2C19 NOS3 MBL2 TNFRSF1A CYP1A1 IL1A ITGA2 F2	Uterus Testis Testis - Leydig cell Testis - germ cell Testis - interstitial cell Testis - seminiferous tubule Salivary gland Trachea Ovary Skin Uterus corpus Tongue Olfactory bulb Retina
DAB2IP	Abdominal Aortic Aneurysm Vascular Infrarenal Myocardial infarction Dilating diathesis	9q33.1 - 9q33.3 (116329225-128995733) → 13Mb	CDKN2A CDKN2B PON1 MMP9 MMP2 MTHFR HMOX1 SERPINE1 PLA2G7 APOE	Whole blood Heart Cardiac myocytes Smooth muscle Atrioventricular node
SPIB	Primary biliary cirrhosis Cholestatic liver disease Anti-mitochondrial antibodies Autoimmune Liver failure Cholestasis Bile duct inflammation	19q13.31 - 19q13.33 (48096232 - 55995612) → 8 Mb	HLA-DPB1 DPB1 HLA-DQA1 CD GSE HLA-DRB1 SS1 DRB1 HLA-DQB1 IDDM1 CELIAC1 IL6 IL12A IL12RB1 IL1RN MBL2 VDR TPMT FAS APT1	CD56+ NK cells CD4+ T cells CD8+ T cells CD19+ B cells Lymph node Fetal liver Liver

			CTLA4 IGF1 KRT19	
MMEL1	See SPIB	1p36.21 - 1p36.33 (1 - 15706544) → 15.7 Mb	See SPIB	See SPIB
TBX2	Heart defects Skeletal malformations Mild mental retardation Growth retardation Cerebellar hypoplasia Interventricular septal defect Patent foramen ovale Aortic coarctation Tricuspid valve insufficiency Mitral valve stenosis	17q23.1 - 17q24.2 (57600001-58300000) → ~10Mb	ACTC1 GATA4 TBX20 MYH6 TLL1 NOTCH1 COL1A1 COL1A2 COL3A1 COL5A2	Heart Cardiac myocytes Smooth muscle Atrioventricular node
RUNX2	Craniosynostosis Bone Metopic synostosis Hypodontia	6p21.1-12.3 (40500001-51800000) → 11.3 MB	FGFR1 FGFR2 FGFR3 TWIST1 MSX2 EFNB1 TGFBR1 TGFBR2 FBN1 RECQL4 RAB23 POR	Bone marrow
CRHR1	Multiple sclerosis Central nervous system Autoimmune disease Inflammation Neurodegeneration	17q21.2 - 17q21.33 (38400001 - 50200000) → 11.8 Mb	HLA-DRB1 IL2RA CLEC16A CD58 TNFRSF1A IRF8 KIF21B TMEM39A HLA-DQB1 CCR5 CD24 CNTF CRYAB IFNG APOE TGFB1 CTLA4 ICAM1 SH2D2A	CD14+ monocytes BDCA4+ Dentritic cells Temporal lobe Globus pallidus Cerebellum peduncles Cerebellum Caudate nucleus Whole brain Parietal lobe Medulla oblongata Amygdala Prefrontal cortex Occipital lobe Hypothalamus Thalamus Subthalamic nucleus Cingulate cortex Fetal brain Adrenal cortex
IFNG	Leprosy Infectious disease Chronic infectious disease Mycobacterium leprae Macrophages Skin Schwann cells Peripheral nerves	12q14.2 - 12q21.1 63100001 - 75700000 → 12.6 Mb	TNF IL10 LTA MRC1 HLA-DRB1 TLR2 PARK2 PACRG NRAMP1	CD56+ NK cells CD4+ T cells CD8+ T cells CD19+ B cells Skin
SH2B1	Renal development Enteric development	16p12.1-11.2 24200001-34600000)	RET PAX2 UPK3A	Fetal liver Liver Kidney

	Kidney Urinary tract	→ 10.4 MB	AGT GATA3 PKD1 PKD2 FCYT HNF1B MCKD1 HNFJ1 NF1 EDNRB EDN1 EDN2 EDN3	Prostate
DISP1	See HCCS	1q41	See HCCS	See HCCS
G6PC3	Congenital neutropenia Pulmonary arterial hypertension Cardiac abnormalities Anemia Marked thymic involution	17q12-q21.31 31800001- 44900000 → 13.1 MB	BMPR2 ALK1 SMAD9 ASD1 GATA4 TBX20 MYH6 ACTC1 TLL1	Whole blood Thymus Lung Heart Fetal lung Cardiac myocytes Atrioventricular node
PQBP1	Mental retardation Microcephaly Short stature Dysmophic craniofacial features Impaired hearing Mild cognitive impairment with difficulty reading and forgetfulness with staring spells Pulmonary nodules Cleft palate Broad distal phalanges of the thumbs and toes	Xp11.3-p11.22 42400001- 54800000 → 12Mb	FLNA ARFGEF2 MEKK4 NAPA LRP2	Temporal lobe Globus pallidus Cerebellum peduncles Cerebellum Caudate nucleus Whole brain Parietal lobe Medulla oblongata Amygdala Prefrontal cortex Occipital lobe Hypothalamus Thalamus Subthalamic nucleus Cingulate cortex Pons Spinal cord Fetal brain Lung Fetal lung
CD320	Methylmalonic acid Metabolic Cobalamin Vitamin B12 Homocystinuria Adenosylcobalamin Methylcobalamin	19p13.2 6899460-12599012 → ~5.7Mb	MMACHC MMAB MUT MTR MMADHC LMBRD1 MMAA SUCLA2 SUCLG1 MCEE	None
CHST14	Ehlers-Danlos syndrome Cutis hyperelastica Connective tissue disorder Collagen Skin hyperextensibility Articular hypermobility Tissue fragility	15q14-15q15 31400001- 42600000 → 11Mb	COL5A1 COL5A2 COL1A1 COL1A2 COL3A1 TNXB PLOD ZNF469 ADAMTS2 GSTP1 GSTM1 B4GALT7	Skin

			TNC TNR	
PLCE1	Esophageal Squamous cell Cancer Adenocarcinoma Dysphagia	10q23 82876191- 96972141 → ~14Mb	TP53 CDKN2A DEC1 DCC DLEC1 TGFBR2 LZTS1 RNF6 WWOX APC ADH1B ALDH2 S100A14 RB1 MCC CRP PTGS2 NQO1 CDKN2A	Colorectal adenocarcinoma Salivary gland
C20orf54	See PLCE1	20p12-p13 1 – 17800000 → ~18Mb	See PLCE1	See PLCE1
SDCCAG8	Nephronophthisis Ciliopathy Dysplasia Kidney Cystic Dyskinesia Retinal degeneration Cilia	1q43-q44 234490661- 247199719 → ~12Mb	OFD1 NPHP1 INVS NPHP3 NPHP4 IQCB1 CEP290 GLIS2 RPGRIP1L NEK8 TMEM67 ARL6 MKKS TTC8 TRIM32 MKS1 C2ORF86 CCDC28B PKD2 PKD1 ALMS1 TMEM216 CC2D2A	Kidney Ciliary ganglion Retina
TP63	Lung adenocarcinoma Lung cancer Carcinoma Neoplasm Lung Cigarette smoking	3q28-q29 189444042- 199446827 → ~10Mb	EGFR TP53 KRAS BRAF ERBB2 MET STK11 PIK3CA PARK2 NKX2-1 DOK2 ERCC6 CHRNA3 CHRNA5 CYP2A6 CASP8 MPO TERT CLPTM1L	Lung Fetal lung Trachea Tongue

			BAT3 APOM OGG1 CYP1A1 ALK EML4	
UBE2E2	Diabetes mellitus type 2 Non-insulin-dependent Diabetes Metabolic disorder Glucose Insulin Maturity-onset diabetes	3p24 16395158-32190491 → ~15Mb	CAPN10 ADIPOQ AKT2 ATP5O CDKN2A CDKN2B COL18A1 ENPP1 EPHX2 FTO G6PC2 SLC2A2 SLC2A4 GPD2 HK1 HNF1A HNF1B HNF4A IL6 IRS1 KCNJ11 KCNJ15 LEPR LIPC LPIN1 MAPK8IP1 NAMPT NEUROD1 NOS1AP PAX4 PPARG PPARGC1A PPP1R3A PSMA6 PTPN1 RETN RUNX2 SGK1 TCF7L2 TNMD UCP2	Whole blood Pancreas Pancreatic islets
LPP	Cardiac anomalies Heart defect Rib anomalies Hypospadias Small kidneys Esophageal Atresia	3q27.1 - q28 182700001-192300000 → ~ 9.5 MB	JAG1 NKX2-5 ZFPM2 GDF1 TBX1 GATA4 NODAL CFC1 ALDH1A2 TDGF1 GATA6	Heart Kidney Cardiac myocytes Atrioventricular node
RANBP1	Eye Eye movement Pursuit eye Movement Schizophrenia	22q11.21 -11.23 17900001-25900000 → ~8MB	ATXN2 ATXN3 INPP5E SETX NYS4 TMEM216 CACNA1A FRMD7	Retina

			PRNP	
HTR7	Alcoholism Event-related brain oscillations	10q23.2 - 23.33 87900001-97000000 → ~9MB	GABRA2 ADH1B TAS2R16 RCBTB1 HTR2A	Temporal lobe Globus pallidus Cerebellum peduncles Cerebellum Caudate nucleus Whole brain Parietal lobe Medulla oblongata Amygdala Prefrontal cortex Occipital lobe Hypothalamus Thalamus Subthalamic nucleus Cingulate cortex Fetal brain Adrenal cortex
SOX17	Kidney Urinary tract Vesico-ureteric Reflux Nephron epithelia CAKUT VUR	8q11.21 - 8q12.3 48Mb - 66Mb → 18Mb	PAX2 PAX8 RET GDNF IL8 VEGFA TGFB1 KLK1 UPK1A UPK3 ACE AGTR1 AGTR2 ROBO2 WNT9B WNT2B WNT11 WNT4 FGF8 FGF9 FGFR3 FGF10 HGFAC EGF	Kidney
ACAD9	Complex I Mitochondria Respiratory chain Coenzyme Q Oxidoreductase Dehydrogenase	3q21.1 - 3q22.3 123Mb - 140Mb → 17Mb	NDUFV1 NDUFV2 NDUFS1 NDUFS2 NDUFS3 NDUFS4 NDUFS6 NDUFS7 NDUFS8 NDUFA2 NDUFA11 NDUFAF3 NDUFAF2 NDUFAF4 C20ORF7 NUBPL FOXRED1 NDUFA1 ND1 ND2 ND3 ND4 ND5 ND6	None

			TRNS2 C8ORF38 SDHA BCS1L COX3 COX10 COX15 SCO2 SURF1 TACO1 MTATP6 TRNV TRNK TRNW TRNL1 DLD PDHA1 LRPPRC EIF2C2	
TRAF3IP2	Psoriasis Psoriasis vulgaris Psoriatic arthritis Immune Inflammation Skin Hyperproliferative	6q21 105Mb - 114Mb → 9Mb	CDKAL1 PTPN22 ADAM33 HLA-C HCR HLA-B CDSN TNF IL1B IL10 IL1RA DEFB104A DEFB4A DEFB104B STAT4 IL13 TAP1 SLC9A3R1 SERPINA1 RAGE MMP2 KIR2DL2 IRF2 IL1RN IL12B HLA-DQB1 HLA-DQA1 HLA-DMB VEGFA HLA-DRB1 KIR2DS1 MIF MICA MICB VDR IL23R PSMB8 PSMB9 IL15 TAP2 TNIP1 APOE COG6 TSC1 IL23A LCE3A LCE3D	Skin

			SPATA2 STAT2 TNFAIP3	
WDR62	Microcephaly Brain Neural Cerebral cortical Mental retardation Developmental Delay	19q12 - 19q13.13 30Mb - 43Mb → 13Mb	MCPH1 CDK5RAP2 CEP152 ASPM CENPJ STIL	Temporal lobe Globus pallidus Cerebellum peduncles Cerebellum Caudate nucleus Whole brain Parietal lobe Medulla oblongata Amygdala Prefrontal cortex Occipital lobe Hypothalamus Thalamus Subthalamic nucleus Cingulate cortex Fetal brain Adrenal cortex

SUPPLEMENTARY TABLE S4: COMPARATIVE RESULTS BETWEEN THE CONFIRMED AND UNCONFIRMED DISEASE-GENE ASSOCIATIONS USING THE CANDIDATE SET BASED PRIORITIZATION TOOLS.

Tools	Suspects		GeneWanderer-RW		Posmed-DN		GeneDistiller		Pinta-CS
	ToppGene		GeneWanderer-DK		Posmed-KS		Endeavour-CS		
Confirmed									
Median	11.43	16.80	24.15	28.89	60.00	33.33	15.79	11.11	20.78
TPR at 5%	17.39	39.13	17.39	8.70	4.35	4.35	17.39	21.74	21.74
TPR at 10%	17.39	47.83	17.39	17.39	8.70	8.70	39.13	39.13	21.74
TPR at 30%	43.48	56.52	65.22	43.48	21.74	26.09	73.91	91.30	65.22
Unconfirmed									
Median	14.12	24.91	13.86	11.63	20	28.13	7.79	11.21	13
TPR at 5%	26.32	31.58	15.79	15.79	5.26	5.26	36.84	31.58	36.84
TPR at 10%	26.32	36.84	36.84	26.32	15.79	5.26	57.89	47.37	42.11
TPR at 30%	36.84	47.37	57.89	63.16	26.32	21.05	84.21	89.47	78.95

SUPPLEMENTARY TABLE S5: COMPARATIVE RESULTS BETWEEN THE CONFIRMED AND UNCONFIRMED DISEASE-GENE ASSOCIATIONS USING THE GENOME-WIDE PRIORITIZATION TOOLS.

Tools	Pinta-GW	Candid	Endeavour-GW
Confirmed			
Median	21.03	25.2	15.97
TPR at 5%	21.74	8.70	21.74
TPR at 10%	21.74	26.09	30.43
TPR at 30%	69.57	60.87	73.91
Unconfirmed			
Median	12.64	17.7	10.82
TPR at 5%	31.58	36.84	36.84
TPR at 10%	42.11	42.11	47.37
TPR at 30%	73.68	68.42	68.42

SUPPLEMENTARY TABLE S6: COMPARATIVE RESULTS BETWEEN THE MONOGENIC DISEASE GENES AND THE COMPLEX DISORDER GENES USING THE CANDIDATE SET BASED GENE PRIORITIZATION TOOLS.

Tools	Suspects		GeneWanderer-RW		Posmed-DN		GeneDistiller		Pinta-CS
	ToppGene		GeneWanderer-DK		Posmed-KS		Endeavour-CS		
Monogenic diseases									
Median	22.76	6.54	21.11	35.94	35.00	27.22	8.08	8.89	20.63
TPR at 5%	11.76	47.06	11.76	11.76	0.00	5.88	29.41	35.29	29.41
TPR at 10%	11.76	52.94	11.76	11.76	5.88	5.88	52.94	52.94	29.41
TPR at 30%	29.41	64.71	52.94	35.29	17.65	17.65	82.35	94.12	64.71
Complex disorders									
Median	10.56	20.33	22.11	22.97	21.82	25.83	11.39	19.12	20.78
TPR at 5%	28	28	12	8	8	4	24	12	20
TPR at 10%	28	36	28	16	16	8	48	32	24
TPR at 30%	48	52	64	52	28	32	84	92	72

SUPPLEMENTARY TABLE S7: COMPARATIVE RESULTS BETWEEN THE MONOGENIC DISEASE GENES AND THE COMPLEX DISORDER GENES USING THE GENOME-WIDE GENE PRIORITIZATION TOOLS.

Tools	Pinta-GW	Candid	Endeavour-GW
Monogenic diseases			
Median	17.62	29.98	7.57
TPR at 5%	35.29	23.53	35.29
TPR at 10%	35.29	29.41	52.94
TPR at 30%	70.59	52.94	82.35
Complex disorders			
Median	19.74	14.20	16.54
TPR at 5%	20	20	24
TPR at 10%	28	36	28
TPR at 30%	72	72	64

SUPPLEMENTARY TABLE S8: COMPARATIVE RESULTS BETWEEN GENE PRIORITIZATION TOOLS TRAINED WITH KNOWN GENES AND DESCRIPTIVE KEYWORDS.

Tools	Suspects		Posmed-DN		GeneWanderer-RW		Endeavour-CS		Pinta-CS		ToppGene	
	Candid		Posmed-KS		GeneWanderer-DK		Endeavour-GW		Pinta-GW		GeneDistiller	
	Keywords				Genes							
Median	12.77	18.10	45.45	31.44	22.11	22.97	11.16	15.49	18.87	19.03	16.80	11.11
Response rate	88.89	100	50	47.62	95.24	88.10	100	100	100	100	97.62	97.62
TPR at 5%	33.33	21.43	4.76	4.76	16.67	11.90	26.19	28.57	28.57	26.19	35.71	26.19
TPR at 10%	33.33	33.33	11.90	7.14	26.19	21.43	42.86	38.10	30.95	30.95	42.86	47.62
TPR at 30%	62.96	64.29	23.81	23.81	61.90	52.38	90.48	71.43	71.43	71.43	52.38	78.57

Average		
	Keywords	Genes
Median	26.94	17.19
Response rate	71.63	97.32
TPR at 5%	16.07	25.00
TPR at 10%	21.43	35.12
TPR at 30%	43.72	68.75

SUPPLEMENTARY TABLE S9: RANKING POSITIONS (AS RANK RATIOS) OF THE 42 NOVEL DISEASE GENES FROM THE VALIDATION DATA SET FOR THE CANDIDATE SET GENE PRIORITIZATION METHODS.

	ToppGene		GeneWanderer-DK		Posmed-KS		Endeavour-CS		
	Suspects		GeneWanderer-RW		Posmed-DN		GeneDistiller		Pinta-CS
HCCS	23.08	3.39	46.81	37.5	15.38	n.a.	7.69	8.89	15.69
BRCA2	63.64	2.13	9.68	8.33	30	37.5	1.28	2.9	2.86
TNFRSF19	14.12	13.04	31.48	41.86	7.69	28.13	31.29	22.76	26.23
MECOM	n.a.	n.a.	45	11.11	n.a.	75	24.34	8.06	26.92
ATF7IP	69.9	37.44	76.47	39.52	n.a.	n.a.	11.39	41.73	66.88
DMRT1	0.97	71.7	28.57	n.a.	21.43	21.43	15.79	97.78	37.5
FUT2	10.48	89.71	50	92.09	n.a.	n.a.	26.19	27.41	17.11
CSF1R	0.94	1.6	6.35	3.77	45.45	15.56	1.12	5	9.76
GLI3	22.06	1.12	2	2.44	50	33.33	0.85	3.85	1.92
STOM	17.65	67.37	6.85	4.69	n.a.	n.a.	35.65	12.71	15.79
UTRN	2.5	2.94	12.5	10.53	20	n.a.	5	2.04	3.33
GABRR1	11.43	5.26	56	54.55	85.71	100	4.55	25.76	12.9
UBE2L3	79.71	87.35	16.83	16.47	n.a.	n.a.	9.14	28.99	2.68
BCL3	0.84	2.17	3.65	11.17	n.a.	n.a.	5.78	6.52	10.58
EZH2	37.37	77.01	18.37	36.59	60.8	21.11	8.05	16.06	20.63
TRAF6	4.48	40.85	6.52	11.63	4.76	23.53	0.82	27.55	4.84
IL10	1.81	1.35	0.87	27.55	2.7	2.56	8.44	28.5	0.5
DAB2IP	49.28	91.25	22	30.23	n.a.	n.a.	37.7	20.34	20.78
SPIB	71.88	16.8	24.73	12.58	94	82.35	6.37	7.69	35.89
MMEL1	4.29	58.74	n.a.	n.a.	n.a.	n.a.	51.71	22.68	24.09
TBX2	n.a.	1.11	n.a.	n.a.	n.a.	n.a.	11.3	0.51	2.86
RUNX2	2.34	1.65	1.14	2.94	9.52	4.26	1.31	2.01	0.99
CRHR1	10.56	3.48	23.58	24.14	21.82	21.31	17.65	10.22	15.18
IFNG	1.56	1.64	2.04	7.89	5.88	10	0.91	2.94	1.92
SH2B1	n.a.	80	53.09	20.55	n.a.	n.a.	7.47	11.81	13
DISP1	67.74	10.87	11.54	n.a.	n.a.	n.a.	8.11	22.22	96.67
G6PC3	22.76	23.86	40.37	n.a.	n.a.	n.a.	11.99	19.12	59.39
PQBP1	offline	6.54	15.22	22.97	n.a.	n.a.	1.14	8.55	38.24
CD320	offline	51.59	74	85.88	n.a.	n.a.	24	23.68	25
CHST14	offline	30.11	20	97.18	100	100	25	7.65	22.22
PLCE1	offline	3.23	22.22	35.29	60	29.55	27.07	11.11	50
C20orf54	offline	36.77	98.88	95.89	n.a.	n.a.	n.a.	97.19	94.06
SDCCAG8	offline	1.64	63.64	93.1	n.a.	n.a.	40	1.3	2.44
TP63	offline	3.23	14.29	15.15	n.a.	n.a.	0.89	1.82	11.63
UBE2E2	offline	94.87	51.85	52.17	n.a.	n.a.	41.56	27.42	75.76
LPP	offline	85.37	61.4	78.26	87.5	40	11.11	11.21	22.73
RANBP1	offline	93.39	28.17	23.73	n.a.	n.a.	18.31	57.93	40.77
HTR7	offline	82.56	1.75	6.38	n.a.	n.a.	0.85	0.89	3.23
SOX17	offline	31.18	2.56	20.59	60	48	5.04	10.2	3.13
ACAD9	offline	1.39	10.89	2.3	68	95.16	31.5	1.58	31.3
TRAF3IP2	offline	8.47	25	6.45	90.91	52.63	14.29	19.33	28.26
WDR62	offline	91.54	95.71	86.67	n.a.	n.a.	37.43	7.14	63.95

SUPPLEMENTARY TABLE S10: RANKING POSITIONS (AS RANK RATIOS) OF THE 42 NOVEL DISEASE GENES FROM THE VALIDATION DATA SET FOR THE GENOME-WIDE PRIORITIZATION METHODS.

	Candid	Endeavour-GW	Pinta-GW
HCCS	25.85	2.75	10.78
BRCA2	1.37	1.25	0.29
TNFRSF19	5.74	36.09	21.24
MECOM	11.6	15.89	30.24
ATF7IP	1.38	49.88	58.21
DMRT1	41.83	79.04	29.94
FUT2	47.67	42.43	19.74
CSF1R	2.93	1.67	2.69
GLI3	7.92	2.25	0.76
STOM	26.54	6.52	22.72
UTRN	18.51	0.46	2.99
GABRR1	17.43	15.97	11.75
UBE2L3	49.17	71.21	6.34
BCL3	60.27	4.74	10.76
EZH2	4.66	34.43	23.41
TRAF6	0.13	16.54	8.11
IL10	13.22	26.26	0.18
DAB2IP	14.2	26.4	21.03
SPIB	5.56	10.28	30.44
MMEL1	25.2	39.37	18.32
TBX2	31.34	1.87	1.34
RUNX2	1.79	3.07	0.18
CRHR1	15.04	13.35	12.94
IFNG	47.19	1.21	0.21
SH2B1	3.53	10.82	12.64
DISP1	2.26	32.85	93.23
G6PC3	47.72	18.72	51.22
PQBP1	22.51	16.74	34.79
CD320	80.54	46.95	26.41
CHST14	61.64	7.57	25.8
PLCE1	10.45	13.03	42.6
C20orf54	67.5	95.54	95.04
SDCCAG8	78.11	5.23	0.85
TP63	7.04	1.35	11.67
UBE2E2	51.18	17.38	61.89
LPP	74.55	6.93	17.62
RANBP1	7.32	46.07	48.2
HTR7	17.7	0.56	1.73
SOX17	74.67	15.08	1.11
ACAD9	33.45	1.53	31.53
TRAF3IP2	2.05	39.04	23.1
WDR62	29.98	23.35	63.97

SUPPLEMENTARY TABLE S11: RANKING POSITIONS (AS RANK RATIOS) OF THE 42 NOVEL DISEASE GENES FROM THE VALIDATION DATA SET FOR THE META PREDICTORS.

	Candidate-set based meta-predictor	Genome-wide meta-predictor
HCCS	5.49	25.59
BRCA2	1.28	0.32
TNFRSF19	10.88	10.99
MECOM	8.21	18.82
ATF7IP	36.19	9.79
DMRT1	29.47	33.41
FUT2	62.36	33.22
CSF1R	2.14	2.04
GLI3	0.85	2.38
STOM	11.02	20.22
UTRN	1.67	6.6
GABRR1	7.58	10.86
UBE2L3	5.65	16.59
BCL3	0.72	28.31
EZH2	15.03	11.03
TRAF6	1.64	0.82
IL10	0.89	1.55
DAB2IP	19.67	14.5
SPIB	5.05	12.28
MMEL1	27.96	17.33
TBX2	0.24	6.1
RUNX2	0.65	0.65
CRHR1	6.41	10.52
IFNG	0.91	3.09
SH2B1	12.81	6.24
DISP1	9.46	12.63
G6PC3	21.34	43.72
PQBP1	2.27	24.34
CD320	39.56	53.81
CHST14	14.75	42.89
PLCE1	10.53	21.9
C20orf54	149.44	63.93
SDCCAG8	3.9	7.58
TP63	0.89	8.16
UBE2E2	102.6	56.12
LPP	28.21	44.43
RANBP1	65.12	18.58
HTR7	0.85	2.74
SOX17	0.84	9
ACAD9	2	27.08
TRAF3IP2	3.36	7.07
WDR62	80.29	49.74

SUPPLEMENTARY TABLE S12: COMPARISON OF THE STATED PERFORMANCE IN THE ORIGINAL TOOL PUBLICATION WITH MEASURED PERFORMANCE IN THIS STUDY.

	Original benchmark			Current study
Tools	**Size (#genes)**	**AUC**	**Reference**	**AUC**
Candid-GW	73	89.5% (a)	Hutz *et al.*	73%
Endeavour-GW	620	94.35%	Tranchevent *et al.*	79%
Suspects	155	87.07%	Adie *et al.*	76%
ToppGene	693	91.6%	Chen *et al.*	66%
	150	92.61% (b)		
GeneWanderer-RW	783	98.1%	Koehler *et al.*	71%
GeneWanderer-DK		> 90.8% (c)		67%
Posmed-DN	N.A.	N.A.	N.A.	56%
Posmed-KS	N.A.	N.A.	N.A.	58%
GeneDistiller	N.A.	N.A.	N.A.	86%
Endeavour-CS	620	94.12%	Tranchevent *et al.*	83%
	12432	91.65%	Schuierer *et al.*	
	11777	92.58%		
Pinta-CS	40	91% (d)	Nitsch *et al.*	75%
Pinta-GW	N.A.	N.A.	N.A.	77%

(a) Reported performance using equal weights for fair comparison.
(b) Non convex AUC
(c) AUC is not given but is higher than 90.8%
(d) Using expression data instead of binay vectors

Chapter 4

Summary

This chapter takes advantage of the information gathered in the two previous chapters to perform a gene prioritization experiment in a real biological experiment with the intention of discovering novel genes and not only validating or reviewing different tools. The top tools found in the previous chapter have been combined in a two-layer based experiment to obtain a ranking of 200 candidate genes in three diseases starting from the whole genome as an initial candidate set.

This work is not the first attempt to compare gene prioritization tools but it is the first time in which performance reasons have been used to select the tools to be combined and that a statistically supported method has been used to combine the rankings of the tools.

Candid [1], Pinta[2], Genedistiller[3] and Endeavour[4], the top four gene prioritization tools from the previous chapter (two genomewide tools and two non-genomewide tools) have been combined in a two steps strategy. First, using the full human genome, the genomewide tools have reduced the list of candidates to 2000 genes (approximately 10%, which according to the previous chapter includes the disease gene around one third of the experiments). The two rankings have been combined using order statistics and a second step including the two non-genomewide tools has been launched and the top 10% has been selected. A final ranking of 200 genes has been obtained using order statistics.

This approach has been applied to three diseases: congenital heart disease, congenital diaphragmatic hernia and asthma. The validation, when using real life data, is much harder since there is not a valid result to compare to. Therefore, the quality of the rankings has been provided by experts in each disease.

In general, a thorough analysis of the top genes shows that genes related to the disease have been highly ranked. After a first analysis of the results, this approach has confirmed the retinoic acid pathway hypothesis for congenital diaphragmatic hernia and has suggested two new candidate genes for asthma: *RELA* and *FAS*. Furthermore, the top ranking gene for congenital heart defect, *BMPR1A* has been recently associated with syndromic heart defects

This work will be submitted to the *Genome Medicine* journal in October 2012

Personal contribution

This chapter has been jointly produced by the Ph.D. candidate and the mentioned co-authors in terms of the conception of the idea and development of the study. In particular,

the Ph.D. candidate has initiated the approach, has run two gene prioritization tools for three diseases, has collaborated in the interpretation of the final ranking related to asthma and has written the manuscript.

Combination of gene prioritization tools gives an insight into disease gene discovery

Francisco Bonachela-Capdevila[1,2,*], Daniela Börnigen[2,3,4*], Léon-Charles Tranchevent[2,3,*], Jeroen Breckpot[5], Paul Brady[5], Bernard Thienpont[5,6], Catherine Laprise[7], Jan Deprest[8,9], Koenraad Devriendt[5], Joris R Vermeesch[5], Bart de Moor[2,3], Yves Moreau[2,3], Patrick De Causmaecker[1,2,§]

[1] CODeS Group, Department of Computer Science, Katholieke Universiteit Leuven, Campus Kortrijk, Belgium

[2] IBBT-K.U.Leuven Future Health Department, Leuven, Belgium

[3] Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium

[4]Biostatistics Department, Harvard School of Public Health, Harvard University, Boston, MA, USA

[5] Center for Human Genetics, Katholieke Universiteit Leuven, Leuven, Belgium

[6] currently at Laboratory of Signalling and Cell Fate & Laboratory of Epigenetics, Babraham Institute, Cambridge, United Kingdom

[7] Département des Sciences Fondamentales, Université du Québec à Chicoutimi, Saguenay, Quebec, Canada

[8] Department of Reproduction, Development and Regeneration, Katholieke Universiteit Leuven, Leuven, Belgium

[9] Center for Surgical Technologies, Katholieke Universiteit Leuven, Leuven, Belgium

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

FBC: Francisco.BonachelaCapdevila@kuleuven-kortrijk.be

DB: dboernig@hsph.harvard.edu

LT: Leon-Charles.Tranchevent@esat.kuleuven.be

JB: jeroen.breckpot@uzleuven.be

PB: Paul.Brady@med.kuleuven.be

BT: bernard.thienpont@babraham.ac.uk

CL: Catherine_Laprise@uqac.ca

JD: jan.deprest@uz.kuleuven.ac.be

KD: koenraad.devriendt@uzleuven.be

JRV: joris.vermeesch@uzleuven.be

BDM: bart.demoor@esat.kuleuven.be

YM: yves.moreau@esat.kuleuven.be

PDC: Patrick.DeCausmaecker@kuleuven-kortrijk.be

Abstract

Background

Genomic approaches to identify the factors underlying genetic diseases often generate lists of candidate genes. In the last decade, several prioritization methods have been developed to help geneticists to identify *in silico* the most promising candidate genes in order to increase the yield of downstream experiments. In this study, we combine the predictions made by four different prioritization tools in order to propose meaningful candidate genes.

Results

We first apply our analysis on two birth defects, non syndromic congenital heart defect and congenital diaphragmatic hernia, and we extend it to asthma, a complex disease. For each disorder, we combine the predictions from four prioritization tools to identify novel candidate genes. *BMPR1A,* the top ranking gene for congenital heart defects has recently been associated with syndromic heart defects. Furthermore, the results confirm the retinoic acid pathway hypothesis for congenital diaphragmatic hernia and propose two new candidate genes for asthma: *RELA* and *FAS*.

Conclusions

Our computational analysis reveals that the top predictions are either associated with syndromic cases of the diseases under study or are functionally linked to previously identified genes and therefore represent potentially new candidate genes. Altogether, our results demonstrate that computational disease gene prioritization methods can be used to quickly retrieve a small set of relevant candidate genes.

Background

Identifying genes associated with Mendelian or complex disorders is a primary aim in human genetics since more effective treatments can be developed from a better understanding of the molecular factors underlying these genetic disorders. Often this gene identifying process begins with a high-throughput experiment, such as an association study, or a transcriptome profiling study that generate a pool of candidate genes (*e.g.*, a large chromosomal region for a specific phenotype). This set of candidate genes is then scrutinized, to identify the most promising candidates in order to increase the yield of the downstream validation. This task can be time-consuming when done manually because a large amount of data needs to be analysed and integrated. However, geneticists can nowadays consider computational predictions as an extra line of evidence, in order to derive a manageable and meaningful set of promising candidate genes. We have, for example, combined copy number variant detection and prioritization to identify *TAB2*

as a novel congenital heart defects gene [5]. More recently, Erlich *et al.* have identified a *KIF1A* mutation in a familial case of spastic paraparesis by combining homozygosity mapping and computational prioritization using multiple tools [6].

Most existing prioritization methods integrate multiple genomic data in order to derive accurate predictions. They however differ by the training information they require, the data and the strategy they use, and the output they produce [7]. Prioritization methods are usually benchmarked on known data [8, 9], which can serve as an estimate of their performance on real case studies [10]. In addition, there exist predictive studies that focus on one specific disorder or phenotype and propose meaningful candidate genes by using several prioritization methods in parallel [11–13]. For instance, starting from a list of 9556 potential candidate genes, Tiffin *et al.* have used seven prioritization methods to identify nine promising candidate genes for type 2 diabetes and five for obesity, including *PGM1* that has since then been associated with type 1 diabetes [14]. A few months later, Elbers *et al.* also reported a study on type 2 diabetes and obesity in which they pinpointed 27 functional candidate genes from an initial set of 612 genes using six prioritization methods. Polymorphisms in one of the predicted genes, *ESR1*, have since then been associated with severe obesity [15] and increased HDL [16]. Another example is the work of Teber *et al.*, who analysed the 9556 candidate genes gathered by Tiffin *et al.* using eight prioritization strategies. They showed that 8 of the 11 genes identified in genome-wide association studies were at the same time also tagged as promising by the computational methods. More recently, we have benchmarked several prioritization tools and showed that the best results are obtained when the rankings provided by the tools are combined (Börningen *et al.*, in review). A main conclusion of these and other studies is that combining several methods usually leads to more accurate predictions [17].

However, the previous attempts of combining gene prioritization methods neither provided any performance reasons to select the tools to be combined nor performed the gene rankings combination with any type of statistical support. In the present study, we aim at a step beyond in gene prioritization rankings combination by using four gene prioritization methods proven to outperform other tools in a benchmark (Börnigen *et al.*, in review), by applying Order Statistics to combine gene rankings and by using the whole human genome as starting point. The four gene prioritization methods (Candid [1], Pinta[2], GeneDistiller[3] and Endeavour[4]) are applied in a two-layer holistic strategy that takes advantage of the most suitable input conditions for every tool (whole human genome for Candid and Pinta and smaller candidate sets for GeneDistiller and Endeavour). Intermediate rankings of every tool have been combined using Order Statistics. This strategy has been applied to three genetic disorders and the top ranking genes have been analyzed. We first focus on non syndromic congenital heart defects, that represent our main field of expertise [5, 18]. We then extend the analysis to another birth defect, congenital diaphragmatic hernia, and to a more complex disorder, asthma.

Results and discussion

In the current study, we aim at defining a manageable set of high-quality candidate genes for congenital heart defects by prioritizing the complete human genome. To this end, we first identify which genes are already associated with heart defects and define keywords that describe this disorder. We then use our two-layer workflow in order to determine which genes represent promising candidate genes for further investigation. Eventually, we extend the analysis to another birth defect (diaphragmatic hernia) and to a complex trait (asthma).

Prioritization using four tools

We have defined a two layer workflow to combine the predictions of the four gene prioritization tools (Figure 1). These tools are chosen for their empirical performance on a recent benchmark (see Methods). For each layer, the results are combined using Order Statistics, which allows genes ranked by only one of the methods to still be included in the combined ranking. Our two layer approach is suitable to prioritize the complete genome. If predictions would be combined in a single layer, the top 10% candidates genes would still represent 2,000 genes, which is too large to be further analysed. Using our two-layer strategy, the number of candidate genes goes from around 20,000 to 2,000 (after first layer), and then to 200 (after second layer). This cut-off of 10% is based on our previous study, which demonstrates that these four prioritization methods are able to rank the disease genes on average in the top 10%.
Also, only a two layer based strategy allows the four selected prioritization tools to be combined since one of them, Genedistiller cannot perform genomewide prioritization.

Case study 1: non syndromic congenital heart defects

Congenital Heart Defects (CHD) are structural malformations of the heart that are present at birth. These defects can involve one or several parts of the heart, such as walls, valves, or incoming/exiting blood vessels. There are different types of CHD that range from simple defects with mild symptoms to more complex defects with severe and sometimes life-threatening consequences. CHD are the single most important congenital cause for perinatal mortality and morbidity, affecting close to 1% of newborn babies (8 per 1000) [19, 20]. CHD are induced either by environmental influences [21], by an altered gene dosage or function [22–24], by stochastic factors [25] or by combinations thereof. In the last decades, mutations in several genes have been associated with monogenic CHD, mainly through linkage analysis in large families in which a CHD segregates as an autosomal-dominant trait [26]. However these causative genes only account for a very small fraction (< 1%) of CHD cases [27], thus representing a serious limitation in the genetic counselling of CHD patients and their families and in the elucidation of the pathogenesis of CHD. Approaches such as the one described in the present study can therefore be used to

pinpoint which genes should be investigated further.

As mentioned in the methods section, a homogeneous gene set is required to build the disease model. We therefore focus on genes associated with isolated CHD since genes associated with syndromic cases would introduce noise in the predictions (because they are also associated with extracardiac phenotypes such as cleft lip/palate). Using the CHD knowledge base CHDWiki [18], we define a set of 30 genes associated with non-syndromic CHD. In addition, we have gathered a set of 23 CHD specific keywords. The lists are provided in table 1.

We then run our two layer approach, the complete results are shown in table A.1 in appendix A. In the following sections, we discuss more extensively the results (genes are mentioned together with their position in the final ranking). A close look at the predictions reveals that 17 out of the 25 top ranked candidates are already associated with other cardiac diseases (see Table 4). In particular, several candidate genes are already known to cause syndromic CHD. For example, several mutations in gene *NOTCH2* (8^{th}) were found in patients with Alagille syndrome, which includes cardiac defects such as pulmonary valvar stenosis, tetralogy of Fallot, and peripheral pulmonary arterial stenoses [28]. A similar situation can be observed with *TGFB1* (11^{th}) and *FBN1* (12^{th}), two genes associated with Marfan syndrome, that includes three cardiac phenotypes: mitral valvar prolapse, mitral regurgitation, and ascending aorta dilation [29]. A third example is the candidate gene *BRAF* (28^{th}), for which mutations have been found in patients with Noonan syndrome, Cardiofaciocutaneous syndrome, and Leopard syndrome, which all include pulmonary valvar stenosis and atrial septal defect (ASD) [30–32]. Other candidate genes are involved in other cardiac diseases such as *SCN5A* (4^{th}) and *LMNA* (5^{th}), associated with dilated cardiomyopathy [33–35]. Also, sequence variants in *MYH7* (2^{nd}), *TNNT2* (14^{th}), and *TNNI3* (47^{th}) are associated with left ventricular hypertrophy (LVWT), which in turns is associated with a higher risk of developing stroke, coronary heart disease, heart failure, and various cardiovascular defects [36]. More interesting maybe are the 8 candidate genes with evidence from animal models (also in Table 4). For instance, mutations in *PDGFRB* (17^{th}) lead to various heart malformations in mice [37]. Similar observations were made when *PTEN* (19^{th}) [38], *AKT1* (25^{th}) [39, 40], and *FBN2* (42^{nd}) [41] are altered.

Then, an enrichment analysis reveals that the remaining candidate genes are associated with functions that are highly relevant to CHD. In particular, several Gene Ontology terms are related to heart development: 'Heart development', 'Cardiac muscle tissue development', 'Blood vessel development', 'Vasculature development'. For example, the gene *TGFB2* (44^{th}) promotes cardiac myocyte differentiation from

embryonic stem cells in mice [42]. Another example is *TTN* (60th) that is involved in the contraction of the cardiac muscle [43]. Several genes are active in the adult heart, and are for instance annotated with the terms 'Heart normal bulk heart 3rd' (CAGP), 'Muscle protein' (SwissProt), and 'Muscle contraction' (Gene Ontology) among others. For example, the cardiac calsequestrins *CASQ2* (16th) is expressed specifically and exclusively in the adult heart [44]. So is the connexin gene *GJA5* (37th), that is present in human ventricle and is involved in gap junction formation [45]. Also, when looking at Cancer Genome Anatomy Project (CGAP) data, it can be observed that several other top candidate genes encode parts of the contractile apparatus and are expressed exclusively in heart (*TNNT2* (14th), *MYBPC3* (3rd), *MYL7* (49th)) or mainly in heart (*MYL2* (43rd), *TNNC1* (30th), *ACTA1* (29th)) [46]. Altogether, this data indicates that the predictions of the computational workflow are relevant. Figure 2 summarizes this information in a network view in which best candidates, known disease genes and links between them are displayed.

We then performed a systematic analysis of CHD associated loci, in order to identify genes that rank high genome-wide and that are located within known CHD loci. For instance, a terminal 4p deletion has been found in a patient with Optiz G/BBB syndrome, which includes cardiac defects [47]. The gene *FGFR3* (88th) is located on 4p16 but has not yet been associated with any cardiac phenotype. Another example is the 5p duplication syndrome, for which heart defects are recurrent. More precisely, a 5p11-p13.3 duplication has been detected in a patient with cardiac phenotype [48]. This region harbours the gene *GDNF*, ranked 108th in our final list, although not yet linked to any cardiac phenotype either. CHD are also associated with the recurrent 10q22-q23 deletion syndrome. This region harbours the *BMPR1A* gene that is ranked 1st in our list. Interestingly, we have recently reported a patient with an atrioventricular septal defect who has an intragenic micro-deletion in the *BMPR1A* gene, therefore confirming its role in heart development [49].

We have included the final predictions in our collaborative platform CHDWiki (accessible at http://homes.esat.kuleuven.be/~biouser/chdwiki) so that researchers interested in heart defects can discuss these candidates.

Case study 2: common left-sided congenital diaphragmatic hernia

In order to show that our strategy is applicable to other genetic disorders, we apply it to another birth defect, common left-sided congenital diaphragmatic hernia. Congenital Diaphragmatic Hernia (CDH) is a defect in the closure or development of the diaphragm characterized by either a structural malformation of the diaphragm or by a defect in its muscularization. In the most extreme cases, the diaphragm can be absent. Its incidence is around 1.7 – 5.7 per 10,000 live births [50]. CDH is generally associated with pulmonary hypoplasia and postnatal pulmonary hypertension of

variable severity that account for the high mortality and much of the morbidity in survivors.

CDH is a complex genetic disorder with numerous genes associated with syndromes in which CDH is observed, but few genes directly associated with development of isolated CDH. Furthermore, CDH does not display complete penetrance for the many loci which share an association, or for those syndromes in which CDH is observed. Two main hypotheses are proposed for development of lung hypoplasia and abnormal pulmonary vasculature observed in cases of CDH; one in which a primary genetic defect causes CDH, leading to an abnormal lung growth and development (secondary hit), and another in which a primary genetic defect directly affects both diaphragm and lung development, supported by the overlapping function of certain genes in the development of both organs.

CDH can be anatomically divided into three main subtypes: a posterolateral 'Bochdalek' hernia (70% of cases), an anterior 'Morgagni' type (27%), and a central septum transversum hernia (3%). In addition, the vast majority of hernias are left-sided (85% of cases), whilst the remainder are right-sided (13%) or bilateral (2%) [51–53]. CDH occurs as an isolated defect in around 60% of cases, or as syndromic CDH for the remainder in which additional congenital malformations are present [54]. The main abnormalities observed in syndromic CDH happen in the cardiovascular, urogenital, and musculoskeletal systems [54].

We have decided to focus on isolated posterolateral CDH since it is the most frequent type. For the prioritization, we have only included the genes and keywords associated to isolated 'Bochdalek' CDH (Table 2). The most promising candidate genes for CDH are shown in table A.2 in appendix A.

Similarly to the other case studies, we also perform an enrichment analysis of the top candidate genes. There are genes involved in 'Heart development' (Gene Ontology), and 'Skeletal system development' (Gene Ontology) that represent two of the three main phenotypes associated with syndromic CDH. For instance, mutations in the gene *GJA1* (39^{th}) are found in patients with heart malformations and defects of laterality [55], which is relevant since laterality may be a factor influencing development of left or right side CDH. Another example is the candidate gene *TTN* (164^{th}) that encodes a large abundant muscle protein. It is involved in skeletal muscle assembly [43], as well as in dilated cardiomyopathy [56], which makes TTN an interesting CDH candidate gene. Regarding the skeletal system, the gene *COL1A1* (89^{th}) is considered interesting through its association with osteogenesis imperfecta, a bone disorder [57], and the *COL3A1* gene has been associated with Ehlers Danlos syndrome in which CDH has been observed as part of the phenotype.

So is *IGF1* (199th) that can induce skeletal muscle hypertrophy through activation of the Ca2+/calmodulin-dependent phosphatase calcineurin [58]. Even more interesting is the *BMP2* gene (184th), involved in bone morphogenesis [59], as well as cardiogenesis [60, 61], two common phenotypes in syndromic CDH.

Interestingly, several Gene Ontology terms are related to cell proliferation and apoptosis: 'regulation of cell proliferation', 'regulation of cell death'. It has indeed been hypothesized that CDH could be caused by perturbation of cell proliferation during diaphragm development [62]. For instance, the blockade of *TGFBR2* (13th) in mice mesoderm-derived tissues results in mildly abnormal lung branching and reduced cell proliferation after mid-gestation, accompanied by multiple defects in other organs, including diaphragmatic hernia [63]. Also, *IGF1* (199th) and *CTNNB1* (22nd) have been shown to be involved in smooth muscle cell proliferation in mouse [64, 65]. *CTNNB1* also plays a role in WNT signalling, a candidate pathway for the pathogenesis of CDH, and also a key pathway in lung development [66].

Another interesting annotation category is about steroids, as indicated by Gene Ontology terms such as 'steroid hormone receptor activity', and 'response to steroid hormone stimulus'. Steroids have been used for a long time already to treat CDH in animal models. For instance, Chen *et al.*, observed that estradiol (a member of the steroid family) can promote lung development in rats with CDH [67]. Similar results were found in an ovine CDH model [68], in a rabbit model [69], and in a rat model [70]. Although there seems to be an effect in animal models, it is still debatable whether prenatal corticosteroids benefit fetuses with CDH [71, 72]. Still several candidate genes detected by our method are involved in steroid activity. For instance, *PPARD* (36th) is a member of the steroid hormone receptor superfamily [73]. *BRCA1* (191st) and *BRCA2* (183rd) mRNA levels are coordinately elevated in human breast cancer cells in response to estrogen, another steroid. [74]. Also, there exists cross-talk between the Wnt and the estrogen signaling pathways via functional interaction between *ESR1* (21st) and *CTNNB1* (22nd) [75].

A common hypothesis is that a defect in the Vitamin A / retinoid signaling pathway may influence the pathogenesis of CDH [76–78]. Several studies have indeed shown that double knockouts of retinoic acid receptors result in impaired lung morphogenesis suggesting that retinoids may be involved in the molecular mechanisms of CDH, [79–81]. It has been demonstrated that prenatal retinoic acid treatment in the nitrofen CDH rodent model during the late stage of lung development could correct the expression level of key lung and/or diaphragm developmental genes like *CTGF* [82], *NR2F2*, *FOG2* and *GATA4* [83]. In our results, and given that no retinoic genes were used for training, we observe that several top candidate genes are involved in the retinoic signaling pathway such as *RARG* (5th),

RXRA (14th), and *RXRG* (78th). This therefore seems to confirm the retinoic acid hypothesis.

We then analyse whether known CDH loci contain genes that are highly ranked genome-wide. We have collected a set of 15 CDH loci, represented in Table 5, and filter the genome-wide results to retain only the genes located within these regions. Only 1 of the 15 regions does not harbour any of the top 2000 genes (7%). The results for the other 14 regions are presented in Table 5. For instance, a known locus on chromosome X contains *AR* that ranks 23th on the genome, and another locus on chromosome 6 contains *GJA1*, ranked 39th. Altogether, these results indicate that the predictions are relevant and can be used by geneticists in their research.

Case study 3: asthma

Our approach can also be applied to more complex diseases. Asthma is a chronic respiratory disease which afflicts patients of all ages, causes significant morbidity and mortality and, consequently, generates substantial costs to our society. It is characterized by variable and recurring symptoms including coughing, chest tightness, wheezing and breathlessness, airflow obstruction, bronchial hyperresponsiveness and underlying inflammation [84].

The prevalence of asthma in western countries has considerably raised during recent decades [85] as well as other allergic and autoimmune disorders, such as rhinitis [86], atopic dermatitis [87], multiple sclerosis [88] and insulin-dependent diabetes [89] mellitus among others [90]. Although multifactorial, asthma is mainly attributable to genetic causes with heritability estimated between 35% and 95% [91]. However, according to the estimation from the largest GWAS in asthma, SNPs explain less than 5% of asthma's heritability [92].

Recent years have seen considerable progress in unraveling the contribution of genetic determinants to the susceptibility of developing asthma as well as to the severity of the disease. Scientific literature indicates that more than 300 genes [93] have been associated with asthma or clinical conditions related to asthma, and the number continues to increase. In addition, environmental factors such as ethnicity and sex of the patient, passive and active tobacco smoke exposure, contact with animals, family size and birth order, vaccination, breastfeeding, and air pollution have been shown to influence asthma [91].
Similarly to the previous two case studies, we have selected a group of keywords describing the characteristics of the disease and a set of training genes associated with asthma. Both the list of keywords and genes, as well as the results the prioritization are available in table A.3 and appendix A.

After applying the two layer gene prioritization strategy, we obtain a prioritized list of 200 genes. An enrichment analysis on the ranking reveals that the top prioritized genes are linked to functions directly connected to asthma, such as 'immune response' (Gene Ontology) or 'defense response' (Gene Ontology). Other terms strongly related to inflammation and immune response arise: 'lymphokine' (Swissprot), 'Cytokine-cytokine receptor interaction' (KEGG), 'cytokine activity' (Gene Ontology) or 'regulation of immunoglobulin production' (Gene Ontology).

A total of 10 genes out of the top 25 have been associated with asthma (see Table 6), some of them in several independent studies such as *TGFB1* (1^{st}) a member of the transforming growth factor-beta family of cytokines, involved in proliferation, differentiation, adhesion, migration and other functions in many different cell types [94, 95]. Other confirmed associations in the top 25 include *IL1B* (2^{nd}) [96], *PTPRC* (3^{rd}) [97], *TGFBR2* (4^{th}) [98], *TLR4* (5^{th}) [99], *IL6* (7^{th}) [100], *TLR2* (10^{th}) [101], *CCL2* (17^{th}) [102, 103] , *IL12B* (23^{th}) [104, 105] and *C3* (24^{th}) [106].

Other 11 genes of the top 25, although not positively associated with asthma, have been related to the disease and to inflammation or remodeling processes in a number of publications and are involved in defense and immune response or interact with other genes that participate in defense and immune response (see Table 6).

Out of these genes, *RELA*, that is involved in induction of pro inflammatory cytokines like TNF and IL1B which are relevant in pathophysiology of asthma and *FAS*, whose soluble portion is different in the context of asthma and rhinitis seem particularly interesting candidates to focus when looking for new genes related to asthma. Besides, *RELA* lies in a known asthma locus, 11q13.1 [107], so is *NFKB1*, on 4q24 [108] and *TNFRSF1A* (26^{th}) in 12p13 [109].

Conclusions

In this study, we propose candidate genes for three genetic disorders, non syndromic congenital heart defects, common left-sided congenital diaphragmatic hernia, and asthma. We use four high performance computational methods and a two layer prioritization strategy to prioritize the complete human genome in order to identify these relevant candidate genes. A preliminary analysis of the results indicates that indeed the top predictions are supported by disease, functional, and animal model evidence. In particular, the role of the top ranking gene for heart defects, *BMP1RA*, has recently been confirmed and among others, *RELA* and *FAS* are suggested as interesting candidate genes for asthma. These promising candidates represent *in silico* predictions that are available to the scientific community for further experimental validation. We believe that such studies are made easier when disease specific knowledge bases exist and are maintained. For heart defects, we use CHDWiki to quickly retrieve the already known genes and to discuss the novel candidate genes.

Methods

Gene prioritization methods

Four gene prioritization tools are considered: Candid [1], Pinta [2], Endeavour [4] and GeneDistiller [3]. We have recently developed a benchmark in the spirit of the CASP challenge (Börnigen *et al.*, in preparation) and have therefore selected these tools on the basis of their respective performance.

Candid accepts keywords that describe the disease under study as input and prioritizes the whole human genome. It mines literature and protein domain databases with the keywords to retrieve potentially relevant genes. Candid further refines the scores by taking into consideration conservation from HomoloGene, and interaction data from EntrezGene, therefore favouring well conserved genes that interact with genes linked to the inputted keywords. Candid version 6 is used, and the four criterions are given equal weights (literature, domains, conservation, and interactions). Pinta maps disease specific differential expression onto a functional network and prioritizes genes according to the differential expression of their functional neighbourhood. In our case, we replace the differential expression data by binary data representing the disease status of the genes. Candidate genes are therefore prioritized according to the number of known disease genes present in their functional neighbourhood. In our case, the functional network is STRING 8.2, the algorithm is heat kernel with diffusion parameter set to 0.5, number of steps to 2 and number of randomizations to 500. Endeavour and GeneDistiller both use a set of known disease genes to model the disease under study. They then rely on a guilt-by-association approach, so that promising candidate genes are the ones that are similar to the already known genes. Both tools use many data types to infer these similarities including functional annotations, literature data, phenotypic data, and regulatory information. For GeneDistiller, the “focus on possible pathways” mode was used. For Endeavour, all data sources except cis-regulatory modules, Bind, IntAct, and Mint were selected.

Disease associated keywords and genes

We have selected three genetic diseases to validate our approach: non-syndromic congenital heart defects, non-syndromic congenital diaphragmatic hernia and asthma.
Identifying appropriate training genes and keywords is crucial in order to retrieve reliable prioritization results. Ideally, gene sets should only contain genes directly associated with the disease of interest, and not weakly connected genes such as biomarkers since homogeneity is very important [8]. In addition, the proportion of genes linked to syndromic cases should be kept under control. Indeed, since these genes are also linked to other phenotypes, using too many of them can introduce

noise in the model and can therefore bias the analysis. Similarly, keywords should be selected wisely. For instance, keywords that represent broad phenotypes (*e.g.*, cancer) should be avoided because they might also introduce noise in the analysis. At contrary, specialized keywords (*e.g.*, oesophageal squamous cell carcinoma) are preferred. Also, when the disease under study involves several phenotypes or known complications, the keyword set should try to cover them, but the focus should always be the main phenotype. For this study, reputed experts have selected, for each of the three diseases, the corresponding sets of genes and keywords. When applicable, several gene and keyword sets were defined for a single disease, but only one was retained after quality control (data not shown). The keywords and training genes that we have collected for the three diseases are presented in tables 1, 2 and 3.

Combination of the predictions

Because we use four prioritization tools, we obtain four different rankings of the candidate genes. We then integrate these four rankings into one global ranking through a two-layer holistic approach (Figure 1). This strategy is based on the results of a recent benchmark, where gene prioritization tools where compared in two different experiments. In the first one, the full genome was prioritized, while for the second experiment a ~10Mb chromosomal region around the true disease gene was used to define the candidate genes (Börnigen *et al.*, in review). We have selected the best tools in each setup as to take advantage of their condition specific performance (Pinta and Candid for full genome, GeneDistiller and Endeavour for non-genomewide). The first layer, then, is designed to filter the whole human genome to identify potentially interesting candidate genes using Candid and Pinta. This layer reduces the search space from approximately 20,000 human genes to a set of 2,000 candidate genes. The second layer is then responsible for selecting the most promising candidate genes from the genes selected in the first layer using Endeavour and GeneDistiller. Again, the top 10% of the gene list is retained, meaning that the final ranking contains 200 candidate genes. After each layer, the rankings are merged using Order Statistics, as used within Endeavour [4]. In both cases and when necessary, gene identifier mapping happens through BioMart [110].

Enrichment analysis

The enrichment analysis is performed using DAVID with the default parameters [111]. The EST data is retrieved from the website of the Cancer Genome Anatomy Project [112]. The functional networks are built with String [113] using knowledge bases and experimental databases only (therefore excluding text-mining) and retaining only high confidence scores (>0.7).

Competing interests

The authors declare no competing interest.

Authors' contributions

The work presented here has been carried out in collaboration between all authors. FBC and LT designed the approach, performed the experiments and wrote the manuscript. JB, BT and KD provided the inputs for congenital heart defects and have analyzed the corresponding predictions. PB, JD and JV provided the inputs for congenital diaphragmatic hernia and have analyzed the associated results. CL provided the keywords and training genes for asthma and analyzed the results. BDM, YV and PDCprovided general guidance. All the authors have revised the manuscript.

Acknowledgements

We kindly thank Sylvain Brohée for helping with the CHDWiki update. We deeply appreciate the assistance of Markus Schülke and Dominik Seelow with GeneDistiller, as well as the help provided by Janna Hutz about the use of Candid. This research was supported by ProMeta, GOA MaNet, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys, START 1, the Flemish Government (FWO: G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), G.0871.12N (Neural circuits) research communities (ICCoS, ANMMM, MLDM), G.0733.09 (3UTR), G.082409 (EGFR); IWT: 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. Hutz JE, Kraja AT, McLeod, HL, Province MA: **CANDID: a flexible method for prioritizing candidate genes for complex human traits**. *Genet. Epidemiol.* 2008, **32**:779–790.

2. Nitsch D, Tranchevent L-C, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y: **Network analysis of differential expression for the identification of disease-causing genes**. *PLoS ONE* 2009, **4**:e5526.

3. Seelow D, Schwarz JM, Schuelke M: **GeneDistiller—Distilling Candidate Genes from Linkage Intervals**. *PLoS ONE* 2008, **3**:e3874.

4. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y: **Gene prioritization through genomic data fusion**. *Nat Biotech* 2006, **24**:537–544.

5. Thienpont B, Zhang L, Postma AV, Breckpot J, Tranchevent L-C, Van Loo P, Møllgård K, Tommerup N, Bache I, Tümer Z, van Engelen K, Menten B, Mortier G, Waggoner D, Gewillig M, Moreau Y, Devriendt K, Larsen LA: **Haploinsufficiency of TAB2 causes congenital heart defects in humans**. *Am. J. Hum. Genet* 2010, **86**:839–849.

6. Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, Dor T, Hannon GJ, Elpeleg O: **Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis**. *Genome Res.* 2011, **21**:658–664.

7. Tranchevent L-C, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y: **A guide to web tools to prioritize candidate genes**. *Brief Bioinform* 2010:bbq007.

8. Schuierer S, Tranchevent L-C, Dengler U, Moreau Y: **Large-scale benchmark of Endeavour using MetaCore maps**. *Bioinformatics* 2010, **26**:1922–1923.

9. Kohler S, Bauer S, Horn D, Robinson P: **Walking the Interactome for Prioritization of Candidate Disease Genes**. *The American Journal of Human Genetics* 2008, **82**:949–958.

10. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG: **Finding function: evaluation methods for functional genomic data**. *BMC Genomics* 2006, **7**:187.

11. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, Lopez-Bigas N, Ouzounis C, Perez-Iratxeta C, Andrade-Navarro MA, Adeyemo A, Patti ME, Semple CAM, Hide W: **Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes**. *Nucleic Acids Res* 2006, **34**:3067–3081.

12. Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, Wijmenga C: **A strategy to search for common obesity and type 2 diabetes genes**. *Trends Endocrinol. Metab* 2007, **18**:19–26.

13. Teber ET, Liu JY, Ballouz S, Fatkin D, Wouters MA: **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.

14. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS: **Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes**. *Nat. Genet.* 2009, **41**:703–707.

15. Chen H-H, Lee W-J, Fann CSJ, Bouchard C, Pan W-H: **Severe obesity is associated with novel single nucleotide polymorphisms of the ESR1 and PPARgamma locus in Han Chinese**. *Am. J. Clin. Nutr.* 2009, **90**:255–262.

16. Figtree GA, Grieve SM, Speller B, Geiger M-J, Robinson BG, Channon KM, Ragoussis J, Collins P, Watkins H: **A commonly occurring polymorphism upstream of the estrogen receptor alpha alters transcription and is associated with increased HDL**. *Atherosclerosis* 2008, **199**:354–361.

17. Thornblad TA, Elliott KS, Jowett J, Visscher PM: **Prioritization of positional candidate genes using multiple web-based software tools**. *Twin Res Hum Genet* 2007, **10**:861–870.

18. Barriot R, Breckpot J, Thienpont B, Brohée S, Van Vooren S, Coessens B, Tranchevent L-C, Van Loo P, Gewillig M, Devriendt K, Moreau Y: **Collaboratively charting the gene-to-phenotype network of human congenital heart defects**. *Genome Med* 2010, **2**:16.

19. Hoffman JIE, Kaplan S: **The incidence of congenital heart disease**. *J. Am. Coll. Cardiol.* 2002, **39**:1890–1900.

20. Thom T, Haase N, Rosamond W, Howard VJ, Rumsfeld J, Manolio T, Zheng Z-J, Flegal K, O'Donnell C, Kittner S, Lloyd-Jones D, Goff DC Jr, Hong Y, Adams R, Friday G, Furie K, Gorelick P, Kissela B, Marler J, Meigs J, Roger V, Sidney S, Sorlie P, Steinberger J, Wasserthiel-Smoller S, Wilson M, Wolf P: **Heart disease and stroke statistics--2006 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee**. *Circulation* 2006, **113**:e85–151.

21. Jenkins KJ, Correa A, Feinstein JA, Botto L, Britt AE, Daniels SR, Elixson M, Warnes CA, Webb CL: **Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics**. *Circulation* 2007, **115**:2995–3014.

22. Bruneau BG: **The developmental genetics of congenital heart disease**. *Nature* 2008, **451**:943–948.

23. Calcagni G, Digilio MC, Sarkozy A, Dallapiccola B, Marino B: **Familial recurrence of congenital heart disease: an overview and review of the literature**. *Eur. J. Pediatr.* 2007, **166**:111–116.

24. Manning N, Archer N: **A study to determine the incidence of structural congenital heart disease in monochorionic twins**. *Prenat. Diagn.* 2006, **26**:1062–1064.

25. Kurnit DM, Layton WM, Matthysse S: **Genetics, chance, and morphogenesis**. *Am. J. Hum. Genet.* 1987, **41**:979–995.

26. Pierpont ME, Basson CT, Benson DW Jr, Gelb BD, Giglia TM, Goldmuntz E, McGee G, Sable CA, Srivastava D, Webb CL: **Genetic basis for congenital heart defects: current knowledge: a scientific statement from the American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics**. *Circulation* 2007, **115**:3015–3038.

27. Posch MG, Perrot A, Schmitt K, Mittelhaus S, Esenwein E-M, Stiller B, Geier C, Dietz R, Gessner R, Ozcelik C, Berger F: **Mutations in GATA4, NKX2.5, CRELD1, and BMP4 are infrequently found in patients with congenital cardiac septal defects**. *Am. J. Med. Genet. A* 2008, **146A**:251–253.

28. McDaniell R, Warthen DM, Sanchez-Lara PA, Pai A, Krantz ID, Piccoli DA, Spinner NB: **NOTCH2 mutations cause Alagille syndrome, a heterogeneous disorder of the notch signaling pathway**. *Am. J. Hum. Genet.* 2006, **79**:169–173.

29. Faivre L, Collod-Beroud G, Loeys BL, Child A, Binquet C, Gautier E, Callewaert B, Arbustini E, Mayer K, Arslan-Kirchner M, Kiotsekoglou A, Comeglio P, Marziliano N, Dietz

HC, Halliday D, Beroud C, Bonithon-Kopp C, Claustres M, Muti C, Plauchu H, Robinson PN, Adès LC, Biggin A, Benetts B, Brett M, Holman KJ, De Backer J, Coucke P, Francke U, De Paepe A, Jondeau G, Boileau C: **Effect of mutation type and location on clinical outcome in 1,013 probands with Marfan syndrome or related phenotypes and FBN1 mutations: an international study**. *Am. J. Hum. Genet.* 2007, **81**:454–466.

30. Niihori T, Aoki Y, Narumi Y, Neri G, Cavé H, Verloes A, Okamoto N, Hennekam RCM, Gillessen-Kaesbach G, Wieczorek D, Kavamura MI, Kurosawa K, Ohashi H, Wilson L, Heron D, Bonneau D, Corona G, Kaname T, Naritomi K, Baumann C, Matsumoto N, Kato K, Kure S, Matsubara Y: **Germline KRAS and BRAF mutations in cardio-facio-cutaneous syndrome**. *Nat. Genet.* 2006, **38**:294–296.

31. Sarkozy A, Carta C, Moretti S, Zampino G, Digilio MC, Pantaleoni F, Scioletti AP, Esposito G, Cordeddu V, Lepri F, Petrangeli V, Dentici ML, Mancini GMS, Selicorni A, Rossi C, Mazzanti L, Marino B, Ferrero GB, Silengo MC, Memo L, Stanzial F, Faravelli F, Stuppia L, Puxeddu E, Gelb BD, Dallapiccola B, Tartaglia M: **Germline BRAF mutations in Noonan, LEOPARD, and cardiofaciocutaneous syndromes: molecular diversity and associated phenotypic spectrum**. *Hum. Mutat.* 2009, **30**:695–702.

32. Rodriguez-Viciana P, Tetsu O, Tidyman WE, Estep AL, Conger BA, Cruz MS, McCormick F, Rauen KA: **Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome**. *Science* 2006, **311**:1287–1290.

33. Fatkin D, MacRae C, Sasaki T, Wolff MR, Porcu M, Frenneaux M, Atherton J, Vidaillet HJ Jr, Spudich S, De Girolami U, Seidman JG, Seidman C, Muntoni F, Müehle G, Johnson W, McDonough B: **Missense mutations in the rod domain of the lamin A/C gene as causes of dilated cardiomyopathy and conduction-system disease**. *N. Engl. J. Med.* 1999, **341**:1715–1724.

34. McPherson E, Turner L, Zador I, Reynolds K, Macgregor D, Giampietro PF: **Ovarian failure and dilated cardiomyopathy due to a novel lamin mutation**. *Am. J. Med. Genet. A* 2009, **149A**:567–572.

35. Sébillon P, Bouchier C, Bidot LD, Bonne G, Ahamed K, Charron P, Drouin-Garraud V, Millaire A, Desrumeaux G, Benaïche A, Charniot J-C, Schwartz K, Villard E, Komajda M: **Expanding the phenotype of LMNA mutations in dilated cardiomyopathy and functional consequences of these mutations**. *J. Med. Genet.* 2003, **40**:560–567.

36. Morita H, Larson MG, Barr SC, Vasan RS, O'Donnell CJ, Hirschhorn JN, Levy D, Corey D, Seidman CE, Seidman JG, Benjamin EJ: **Single-gene mutations and increased left ventricular wall thickness in the community: the Framingham Heart Study**. *Circulation* 2006, **113**:2697–2705.

37. Van den Akker NMS, Winkel LCJ, Nisancioglu MH, Maas S, Wisse LJ, Armulik A, Poelmann RE, Lie-Venema H, Betsholtz C, Gittenberger-de Groot AC: **PDGF-B signaling is important for murine cardiac development: its role in developing atrioventricular valves, coronaries, and cardiac innervation**. *Dev. Dyn.* 2008, **237**:494–503.

38. Oudit GY, Kassiri Z, Zhou J, Liu QC, Liu PP, Backx PH, Dawood F, Crackower MA, Scholey JW, Penninger JM: **Loss of PTEN attenuates the development of pathological**

hypertrophy and heart failure in response to biomechanical stress. *Cardiovasc. Res.* 2008, **78**:505–514.

39. Easton RM, Cho H, Roovers K, Shineman DW, Mizrahi M, Forman MS, Lee VM-Y, Szabolcs M, de Jong R, Oltersdorf T, Ludwig T, Efstratiadis A, Birnbaum MJ: **Role for Akt3/protein kinase Bgamma in attainment of normal brain size**. *Mol. Cell. Biol.* 2005, **25**:1869–1878.

40. Nishi J, Minamino T, Miyauchi H, Nojima A, Tateno K, Okada S, Orimo M, Moriya J, Fong G-H, Sunagawa K, Shibuya M, Komuro I: **Vascular endothelial growth factor receptor-1 regulates postnatal angiogenesis through inhibition of the excessive activation of Akt**. *Circ. Res.* 2008, **103**:261–268.

41. Carta L, Pereira L, Arteaga-Solis E, Lee-Arteaga SY, Lenart B, Starcher B, Merkel CA, Sukoyan M, Kerkis A, Hazeki N, Keene DR, Sakai LY, Ramirez F: **Fibrillins 1 and 2 perform partially overlapping functions during aortic development**. *J. Biol. Chem.* 2006, **281**:8016–8023.

42. Singla DK, Kumar D, Sun B: **Transforming growth factor-beta2 enhances differentiation of cardiac myocytes from embryonic stem cells**. *Biochem. Biophys. Res. Commun.* 2005, **332**:135–141.

43. Gregorio CC, Trombitás K, Centner T, Kolmerer B, Stier G, Kunke K, Suzuki K, Obermayr F, Herrmann B, Granzier H, Sorimachi H, Labeit S: **The NH2 terminus of titin spans the Z-disc: its interaction with a novel 19-kD ligand (T-cap) is required for sarcomeric integrity**. *J. Cell Biol.* 1998, **143**:1013–1027.

44. Park KW, Goo JH, Chung HS, Kim H, Kim DH, Park WJ: **Cloning of the genes encoding mouse cardiac and skeletal calsequestrins: expression pattern during embryogenesis**. *Gene* 1998, **217**:25–30.

45. Kanter HL, Saffitz JE, Beyer EC: **Molecular cloning of two human cardiac gap junction proteins, connexin40 and connexin45**. *J. Mol. Cell. Cardiol.* 1994, **26**:861–868.

46. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: **SAGEmap: a public gene expression resource**. *Genome Res.* 2000, **10**:1051–1060.

47. So J, Müller I, Kunath M, Herrmann S, Ullmann R, Schweiger S: **Diagnosis of a terminal deletion of 4p with duplication of Xp22.31 in a patient with findings of Opitz G/BBB syndrome and Wolf-Hirschhorn syndrome**. *Am. J. Med. Genet. A* 2008, **146A**:103–109.

48. Loscalzo ML, Becker TA, Sutcliffe M: **A patient with an interstitial duplication of chromosome 5p11-p13.3 further confirming a critical region for 5p duplication syndrome**. *Eur J Med Genet* 2008, **51**:54–60.

49. Breckpot J, Tranchevent L-C, Thienpont B, Bauters M, Troost E, Gewillig M, Vermeesch JR, Moreau Y, Devriendt K, Van Esch H: **BMPR1A is a candidate gene for congenital heart defects associated with the recurrent 10q22q23 deletion syndrome**. *Eur J Med Genet* 2012, **55**:12–16.

50. Kotecha S, Barbato A, Bush A, Claus F, Davenport M, Delacourt C, Deprest J, Eber E, Frenckner B, Greenough A, Nicholson A, Antón-Pacheco JL, Midulla F: **European respiratory society task force on congenital diaphragmatic hernia**. *Eur. Respir. J.* 2011.

51. Torfs CP, Curry CJ, Bateson TF, Honoré LH: **A population-based study of congenital diaphragmatic hernia**. *Teratology* 1992, **46**:555–565.

52. van Loenhout RB, Tibboel D, Post M, Keijzer R: **Congenital diaphragmatic hernia: comparison of animal models and relevance to the human situation**. *Neonatology* 2009, **96**:137–149.

53. Pober BR: **Overview of epidemiology, genetics, birth defects, and chromosome abnormalities associated with CDH**. *Am J Med Genet C Semin Med Genet* 2007, **145C**:158–171.

54. Stoll C, Alembik Y, Dott B, Roth M-P: **Associated malformations in cases with congenital diaphragmatic hernia**. *Genet. Couns.* 2008, **19**:331–339.

55. Britz-Cunningham SH, Shah MM, Zuppan CW, Fletcher WH: **Mutations of the Connexin43 gap-junction gene in patients with heart malformations and defects of laterality**. *N. Engl. J. Med.* 1995, **332**:1323–1329.

56. Itoh-Satoh M, Hayashi T, Nishi H, Koga Y, Arimura T, Koyanagi T, Takahashi M, Hohda S, Ueda K, Nouchi T, Hiroe M, Marumo F, Imaizumi T, Yasunami M, Kimura A: **Titin mutations as the molecular basis for dilated cardiomyopathy**. *Biochem. Biophys. Res. Commun.* 2002, **291**:385–393.

57. Chamberlain JR, Schwarze U, Wang P-R, Hirata RK, Hankenson KD, Pace JM, Underwood RA, Song KM, Sussman M, Byers PH, Russell DW: **Gene targeting in stem cells from individuals with osteogenesis imperfecta**. *Science* 2004, **303**:1198–1201.

58. Semsarian C, Wu MJ, Ju YK, Marciniec T, Yeoh T, Allen DG, Harvey RP, Graham RM: **Skeletal muscle hypertrophy is mediated by a Ca2+-dependent calcineurin signalling pathway**. *Nature* 1999, **400**:576–581.

59. Wang EA, Rosen V, D'Alessandro JS, Bauduy M, Cordes P, Harada T, Israel DI, Hewick RM, Kerns KM, LaPan P: **Recombinant human bone morphogenetic protein induces bone formation**. *Proc. Natl. Acad. Sci. U.S.A.* 1990, **87**:2220–2224.

60. Sugi Y, Yamamura H, Okagawa H, Markwald RR: **Bone morphogenetic protein-2 can mediate myocardial regulation of atrioventricular cushion mesenchymal cell formation in mice**. *Dev. Biol.* 2004, **269**:505–518.

61. Stefanovic S, Abboud N, Désilets S, Nury D, Cowan C, Pucéat M: **Interplay of Oct4 with Sox2 and Sox17: a molecular switch from stem cell pluripotency to specifying a cardiac fate**. *J. Cell Biol.* 2009, **186**:665–673.

62. Clugston RD, Zhang W, Greer JJ: **Early development of the primordial mammalian diaphragm and cellular mechanisms of nitrofen-induced congenital diaphragmatic hernia**. *Birth Defects Res. Part A Clin. Mol. Teratol.* 2010, **88**:15–24.

63. Chen H, Zhuang F, Liu Y-H, Xu B, Del Moral P, Deng W, Chai Y, Kolb M, Gauldie J, Warburton D, Moses HL, Shi W: **TGF-beta receptor II in epithelia versus mesenchyme plays distinct roles in the developing lung**. *Eur. Respir. J.* 2008, **32**:285–295.

64. Jia G, Cheng G, Agrawal DK: **Differential effects of insulin-like growth factor-1 and atheroma-associated cytokines on cell proliferation and apoptosis in plaque smooth muscle cells of symptomatic and asymptomatic patients with carotid stenosis**. *Immunol. Cell Biol.* 2006, **84**:422–429.

65. Quasnichka H, Slater SC, Beeching CA, Boehm M, Sala-Newby GB, George SJ: **Regulation of smooth muscle cell proliferation by beta-catenin/T-cell factor signaling involves modulation of cyclin D1 and p21 expression**. *Circ. Res.* 2006, **99**:1329–1337.

66. De Langhe SP, Reynolds SD: **Wnt signaling in lung organogenesis**. *Organogenesis* 2008, **4**:100–108.

67. Chen G, Qiao Y, Xiao X, Zheng S, Chen L: **Effects of estrogen on lung development in a rat model of diaphragmatic hernia**. *J. Pediatr. Surg.* 2010, **45**:2340–2345.

68. Paddock H, Beierle EA, Chen MK, Mullett T, Wood CM, Kays DW, Langham MR Jr: **Administration of prenatal betamethasone suppresses the adrenal-hypophyseal axis in newborns with congenital diaphragmatic hernia**. *J. Pediatr. Surg.* 2004, **39**:1176–1182.

69. Roubliova XI, Lewi PJ, Verbeken EK, Vaast P, Jani JC, Lu H, Tibboel D, Deprest JA: **The effect of maternal betamethasone and fetal tracheal occlusion on pulmonary vascular morphometry in fetal rabbits with surgically induced diaphragmatic hernia: a placebo controlled morphologic study**. *Prenat. Diagn.* 2009, **29**:674–681.

70. Mayer S, Klaritsch P, Sbragia L, Toelen J, Till H, Deprest JA: **Maternal administration of betamethasone inhibits proliferation induced by fetal tracheal occlusion in the nitrofen rat model for congenital diaphragmatic hernia: a placebo-controlled study**. *Pediatr. Surg. Int.* 2008, **24**:1287–1295.

71. Lally KP, Bagolan P, Hosie S, Lally PA, Stewart M, Cotten CM, Van Meurs KP, Alexander G: **Corticosteroids for fetuses with congenital diaphragmatic hernia: can we show benefit?** *J. Pediatr. Surg.* 2006, **41**:668–674; discussion 668–674.

72. Braby J: **Current and emerging treatment for congenital diaphragmatic hernia**. *Neonatal Netw* 2001, **20**:5–15.

73. Schmidt A, Endo N, Rutledge SJ, Vogel R, Shinar D, Rodan GA: **Identification of a new member of the steroid hormone receptor superfamily that is activated by a peroxisome proliferator and fatty acids**. *Mol. Endocrinol.* 1992, **6**:1634–1641.

74. Spillman MA, Bowcock AM: **BRCA1 and BRCA2 mRNA levels are coordinately elevated in human breast cancer cells in response to estrogen**. *Oncogene* 1996, **13**:1639–1645.

75. Kouzmenko AP, Takeyama K-I, Ito S, Furutani T, Sawatsubashi S, Maki A, Suzuki E, Kawasaki Y, Akiyama T, Tabata T, Kato S: **Wnt/beta-catenin and estrogen signaling converge in vivo**. *J. Biol. Chem.* 2004, **279**:40255–40258.

76. Clugston RD, Zhang W, Alvarez S, de Lera AR, Greer JJ: **Understanding abnormal retinoid signaling as a causative mechanism in congenital diaphragmatic hernia**. *Am. J. Respir. Cell Mol. Biol.* 2010, **42**:276–285.

77. Goumy C, Gouas L, Marceau G, Coste K, Veronese L, Gallot D, Sapin V, Vago P, Tchirkov A: **Retinoid pathway and congenital diaphragmatic hernia: hypothesis from the analysis of chromosomal abnormalities**. *Fetal. Diagn. Ther.* 2010, **28**:129–139.

78. Montedonico S, Nakazawa N, Puri P: **Congenital diaphragmatic hernia and retinoids: searching for an etiology**. *Pediatr. Surg. Int.* 2008, **24**:755–761.

79. Antipatis C, Ashworth CJ, Grant G, Lea RG, Hay SM, Rees WD: **Effects of maternal vitamin A status on fetal heart and lung: changes in expression of key developmental genes**. *Am. J. Physiol.* 1998, **275**:L1184–1191.

80. Nakazawa N, Takayasu H, Montedonico S, Puri P: **Altered regulation of retinoic acid synthesis in nitrofen-induced hypoplastic lung**. *Pediatr. Surg. Int.* 2007, **23**:391–396.

81. Thébaud B, Tibboel D, Rambaud C, Mercier JC, Bourbon JR, Dinh-Xuan AT, Archer SL: **Vitamin A decreases the incidence and severity of nitrofen-induced congenital diaphragmatic hernia in rats**. *Am. J. Physiol.* 1999, **277**:L423–429.

82. Ruttenstock EM, Doi T, Dingemann J, Puri P: **Prenatal administration of retinoic acid upregulates connective tissue growth factor in the nitrofen CDH model**. *Pediatr. Surg. Int.* 2011, **27**:573–577.

83. Doi T, Sugimoto K, Puri P: **Prenatal retinoic acid up-regulates pulmonary gene expression of COUP-TFII, FOG2, and GATA4 in pulmonary hypoplasia**. *J. Pediatr. Surg.* 2009, **44**:1933–1937.

84. **National Asthma Education and Prevention Program. 2007** [http://www.nhlbi.nih.gov/guidelines/asthma/].

85. Beasley R: **The burden of asthma with specific reference to the United States**. *Journal of Allergy and Clinical Immunology* 2002, **109**:S482–S489.

86. Upton MN, McConnachie A, McSharry C, Hart CL, Smith GD, Gillis CR, Watt GC: **Intergenerational 20 year trends in the prevalence of asthma and hay fever in adults: the Midspan family study surveys of parents and offspring**. *BMJ* 2000, **321**:88–92.

87. Williams HC: **Is the prevalence of atopic dermatitis increasing?** *Clin. Exp. Dermatol.* 1992, **17**:385–391.

88. Poser S, Stickel B, Krtsch U, Burckhardt D, Nordman B: **Increasing incidence of multiple sclerosis in South Lower Saxony, Germany**. *Neuroepidemiology* 1989, **8**:207–213.

89. **Variation and trends in incidence of childhood diabetes in Europe. EURODIAB ACE Study Group**. *Lancet* 2000, **355**:873–876.

90. Bach J-F: **The effect of infections on susceptibility to autoimmune and allergic diseases**. *N. Engl. J. Med* 2002, **347**:911–920.

91. Yao T-C, Ober C: **The genetics of asthma and allergic disease: a 21st century perspective.** *Immunological Reviews* 2011, **242**:10–30.

92. Moffatt M, Gut I, Demenais F, Strachan D, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson W, GABRIEL Consortium: **A large-scale, consortium-based genomewide association study of asthma**. *New England Journal of Medicine* 2010, **363**:1211–21.

93. Lee S-H, Park J-S, Park C-S: **The search for genetic variants and epigenetics related to asthma**. *Allergy Asthma Immunol Res* 2011, **3**:236–244.

94. Pulleyn LJ, Newton R, Adcock IM, Barnes PJ: **TGFbeta1 allele association with asthma severity**. *Hum. Genet.* 2001, **109**:623–627.

95. Nagpal K, Sharma S, B-Rao C, Nahid S, Niphadkar PV, Sharma SK, Ghosh B: **TGFbeta1 haplotypes and asthma in Indian populations**. *J. Allergy Clin. Immunol.* 2005, **115**:527–533.

96. Daley D, Lemire M, Akhabir L, Chan-Yeung M, He JQ, McDonald T, Sandford A, Stefanowicz D, Tripp B, Zamar D, Bosse Y, Ferretti V, Montpetit A, Tessier M-C, Becker A, Kozyrskyj AL, Beilby J, McCaskie PA, Musk B, Warrington N, James A, Laprise C, Palmer LJ, Paré PD, Hudson TJ: **Analyses of associations with asthma in four asthma population samples from Canada and Australia**. *Hum. Genet.* 2009, **125**:445–459.

97. Tremblay K, Lemire M, Potvin C, Tremblay A, Hunninghake GM, Raby BA, Hudson TJ, Perez-Iratxeta C, Andrade-Navarro MA, Laprise C: **Genes to Diseases (G2D) Computational Method to Identify Asthma Candidate Genes**. *PLoS ONE* 2008, **3**:e2907.

98. Bottema RWB, Kerkhof M, Reijmerink NE, Thijs C, Smit HA, van Schayck CP, Brunekreef B, van Oosterhout AJ, Postma DS, Koppelman GH: **Gene-gene interaction in regulatory T-cell function in atopy and asthma development in childhood**. *J. Allergy Clin. Immunol* 2010, **126**:338–346, 346.e1–10.

99. Fageråas Böttcher M, Hmani-Aifa M, Lindström A, Jenmalm MC, Mai X-M, Nilsson L, Zdolsek HA, Björkstén B, Söderkvist P, Vaarala O: **A TLR4 polymorphism is associated with asthma and reduced lipopolysaccharide-induced interleukin-12(p70) responses in Swedish children☆**. *Journal of Allergy and Clinical Immunology* 2004, **114**:561–567.

100. Corvol H, De Giacomo A, Eng C, Seibold M, Ziv E, Chapela R, Rodriguez-Santana JR, Rodriguez-Cintron W, Thyne S, Watson HG, Meade K, LeNoir M, Avila PC, Choudhry S, Burchard EG: **Genetic ancestry modifies pharmacogenetic gene-gene interaction for asthma**. *Pharmacogenet. Genomics* 2009, **19**:489–496.

101. Eder W, Klimecki W, Yu L, von Mutius E, Riedler J, Braun-Fahrländer C, Nowak D, Martinez FD: **Toll-like receptor 2 as a major gene for asthma in children of European farmers**. *J. Allergy Clin. Immunol.* 2004, **113**:482–488.

102. Szalai C, Kozma GT, Nagy A, Bojszkó Á, Krikovszky D, Szabó T, Falus A: **Polymorphism in the gene regulatory region of MCP-1 is associated with asthma susceptibility and severity**. *Journal of Allergy and Clinical Immunology* 2001, **108**:375–381.

103. Chelbi H, Ghadiri A, Lacheb J, Ghandil P, Hamzaoui K, Hamzaoui A, Combadiere C: **A polymorphism in the CCL2 chemokine gene is associated with asthma risk: a case-control and a family study in Tunisia**. *Genes Immun.* 2008, **9**:575–581.

104. Morahan G, Huang D, Wu M, Holt BJ, White GP, Kendall GE, Sly PD, Holt PG: **Association of IL12B promoter polymorphism with severity of atopic and non-atopic asthma in children**. *Lancet* 2002, **360**:455–459.

105. Randolph AG, Lange C, Silverman EK, Lazarus R, Silverman ES, Raby B, Brown A, Ozonoff A, Richter B, Weiss ST: **The IL12B gene is associated with asthma**. *Am. J. Hum. Genet.* 2004, **75**:709–715.

106. Hasegawa K, Tamari M, Shao C, Shimizu M, Takahashi N, Mao X-Q, Yamasaki A, Kamada F, Doi S, Fujiwara H, Miyatake A, Fujita K, Tamura G, Matsubara Y, Shirakawa T, Suzuki Y: **Variations in the C3, C3a receptor, and C5 genes affect susceptibility to bronchial asthma**. *Hum. Genet.* 2004, **115**:295–301.

107. Adra CN, Mao XQ, Kawada H, Gao PS, Korzycka B, Donate JL, Shaldon SR, Coull P, Dubowitz M, Enomoto T, Ozawa A, Syed SA, Horiuchi T, Khaeraja R, Khan R, Lin SR, Flinter F, Beales P, Hagihara A, Inoko H, Shirakawa T, Hopkin JM: **Chromosome 11q13 and atopic asthma**. *Clin. Genet.* 1999, **55**:431–437.

108. Aschard H, Bouzigon E, Corda E, Ulgen A, Dizier M-H, Gormand F, Lathrop M, Kauffmann F, Demenais F: **Sex-specific effect of IL9 polymorphisms on lung function and polysensitization**. *Genes Immun.* 2009, **10**:559–565.

109. Bouzigon E, Dizier M-H, Krähenbühl C, Lemainque A, Annesi-Maesano I, Betard C, Bousquet J, Charpin D, Gormand F, Guilloud-Bataille M, Just J, Moual NL, Maccario J, Matran R, Neukirch F, Oryszczyn M-P, Paty E, Pin I, Rosenberg-Bourgin M, Vervloet D, Kauffmann F, Lathrop M, Demenais F: **Clustering Patterns of LOD Scores for Asthma-Related Phenotypes Revealed by a Genome-Wide Screen in 295 French EGEA Families**. *Hum. Mol. Genet.* 2004, **13**:3103–3113.

110. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: **BioMart Central Portal--unified access to biological data**. *Nucleic Acids Research* 2009, **37**:W23–W27.

111. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**:44–57.

112. Brentani H, Caballero OL, Camargo AA, et al.: **The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags**. *Proc. Natl. Acad. Sci. U.S.A.* 2003, **100**:13418–13423.

113. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored**. *Nucleic Acids Res.* 2011, **39**:D561–568.

114. Wang G-S, Cooper TA: **Splicing in disease: disruption of the splicing code and the decoding machinery**. *Nature Reviews Genetics* 2007, **8**:749–761.

115. Bonnert TP, Garka KE, Parnet P, Sonoda G, Testa JR, Sims JE: **The cloning and characterization of human MyD88: a member of an IL-1 receptor related family**. *FEBS Lett.* 1997, **402**:81–84.

116. Sagara H, Okada T, Okumura K, Ogawa H, Ra C, Fukuda T, Nakao A: **Activation of TGF-beta/Smad2 signaling is associated with airway remodeling in asthma**. *J. Allergy Clin. Immunol.* 2002, **110**:249–254.

117. Drube S, Heink S, Walter S, Löhn T, Grusser M, Gerbaulet A, Berod L, Schons J, Dudeck A, Freitag J, Grotha S, Reich D, Rudeschko O, Norgauer J, Hartmann K, Roers A, Kamradt T: **The receptor tyrosine kinase c-Kit controls IL-33 receptor signaling in mast cells**. *Blood* 2010, **115**:3899–3906.

118. Lockett A, Goebl MG, Harrington MA: **Transient membrane recruitment of IRAK-1 in response to LPS and IL-1beta requires TNF R1**. *Am. J. Physiol., Cell Physiol.* 2008, **295**:C313–323.

119. Thompson C, Cloutier A, Bossé Y, Thivierge M, Gouill CL, Larivée P, McDonald PP, Stankova J, Rola-Pleszczynski M: **CysLT1 receptor engagement induces activator protein-1- and NF-kappaB-dependent IL-8 expression**. *Am. J. Respir. Cell Mol. Biol.* 2006, **35**:697–704.

120. Cekic C, Casella CR, Sag D, Antignano F, Kolb J, Suttles J, Hughes MR, Krystal G, Mitchell TC: **MyD88-dependent SHIP1 regulates proinflammatory signaling pathways in dendritic cells after monophosphoryl lipid A stimulation of TLR4**. *J. Immunol.* 2011, **186**:3858–3865.

121. Bhattacharyya S, Balakathiresan NS, Dalgard C, Gutti U, Armistead D, Jozwik C, Srivastava M, Pollard HB, Biswas R: **Elevated miR-155 promotes inflammation in cystic fibrosis by driving hyperexpression of interleukin-8**. *J. Biol. Chem.* 2011, **286**:11604–11615.

122. Pedersen IM, Otero D, Kao E, Miletic AV, Hother C, Ralfkiaer E, Rickert RC, Gronbaek K, David M: **Onco-miR-155 targets SHIP1 to promote TNFalpha-dependent growth of B cell lymphomas**. *EMBO Mol Med* 2009, **1**:288–295.

123. Robinson DS, Hamid Q, Ying S, Tsicopoulos A, Barkans J, Bentley AM, Corrigan C, Durham SR, Kay AB: **Predominant TH2-like bronchoalveolar T-lymphocyte population in atopic asthma**. *N. Engl. J. Med.* 1992, **326**:298–304.

124. Lim C-A, Yao F, Wong JJ-Y, George J, Xu H, Chiu KP, Sung W-K, Lipovich L, Vega VB, Chen J, Shahab A, Zhao XD, Hibberd M, Wei C-L, Lim B, Ng H-H, Ruan Y, Chin K-C: **Genome-wide mapping of RELA(p65) binding identifies E2F1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation**. *Mol. Cell* 2007, **27**:622–635.

125. Mountz JD, Talal N: **Retroviruses, apoptosis and autogenes**. *Immunol. Today* 1993, **14**:532–536.

126. Ho C-Y, Wong C-K, Ko FW-S, Chan CH-S, Ho AS-S, Hui DS-C, Lam CW-K: **Apoptosis and B-cell lymphoma-2 of peripheral blood T lymphocytes and soluble fas in patients with allergic asthma**. *Chest* 2002, **122**:1751–1758.

127. Harker N, Naito T, Cortes M, Hostert A, Hirschberg S, Tolaini M, Roderick K, Georgopoulos K, Kioussis D: **The CD8alpha gene locus is regulated by the Ikaros family of proteins**. *Mol. Cell* 2002, **10**:1403–1415.

128. Valapour M, Guo J, Schroeder JT, Keen J, Cianferoni A, Casolaro V, Georas SN: **Histone deacetylation inhibits IL4 gene expression in T cells**. *J. Allergy Clin. Immunol.* 2002, **109**:238–245.

129. Günther C, Wozel G, Meurer M, Pfeiffer C: **Up-regulation of CCL11 and CCL26 is associated with activated eosinophils in bullous pemphigoid**. *Clin. Exp. Immunol.* 2011, **166**:145–153.

130. Watkins NA, Gusnanto A, de Bono B, De S, Miranda-Saavedra D, Hardie DL, Angenent WGJ, Attwood AP, Ellis PD, Erber W, Foad NS, Garner SF, Isacke CM, Jolley J, Koch K, Macaulay IC, Morley SL, Rendon A, Rice KM, Taylor N, Thijssen-Timmer DC, Tijssen MR, van der Schoot CE, Wernisch L, Winzer T, Dudbridge F, Buckley CD, Langford CF, Teichmann S, Göttgens B, Ouwehand WH: **A HaemAtlas: characterizing gene expression in differentiated human blood cells**. *Blood* 2009, **113**:e1–9.

131. Steele BM, Harper MT, Macaulay IC, Morrell CN, Perez-Tamayo A, Foy M, Habas R, Poole AW, Fitzgerald DJ, Maguire PB: **Canonical Wnt signaling negatively regulates platelet function**. *Proc. Natl. Acad. Sci. U.S.A.* 2009, **106**:19836–19841.

132. Iseki H, Takeda A, Andoh T, Takahashi N, Kurochkin IV, Yarmishyn A, Shimada H, Okazaki Y, Koyama I: **Human Arm protein lost in epithelial cancers, on chromosome X 1 (ALEX1) gene is transcriptionally regulated by CREB and Wnt/beta-catenin signaling**. *Cancer Sci.* 2010, **101**:1361–1366.

133. Stalker TJ, Wu J, Morgans A, Traxler EA, Wang L, Chatterjee MS, Lee D, Quertermous T, Hall RA, Hammer DA, Diamond SL, Brass LF: **Endothelial cell specific adhesion molecule (ESAM) localizes to platelet-platelet contacts and regulates thrombus formation in vivo**. *J. Thromb. Haemost.* 2009, **7**:1886–1896.

134. O'Connor MN, Salles II, Cvejic A, Watkins NA, Walker A, Garner SF, Jones CI, Macaulay IC, Steward M, Zwaginga J-J, Bray SL, Dudbridge F, de Bono B, Goodall AH, Deckmyn H, Stemple DL, Ouwehand WH: **Functional genomics in zebrafish permits rapid characterization of novel platelet membrane proteins**. *Blood* 2009, **113**:4754–4762.

135. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data**. *Bioinformatics* 2006, **22**:1122–1129.

136. Ben-Dor A, Chor B, Karp R, Yakhini Z: **Discovering local structure in gene expression data: the order-preserving submatrix problem**. *J. Comput. Biol* 2003, **10**:373–384.

137. Ihmels J, Bergmann S, Barkai N: **Defining transcription modules using large-scale gene expression data**. *Bioinformatics* 2004, **20**:1993 –2003.

138. Tanay A, Sharan R, Shamir R: **Discovering statistically significant biclusters in gene expression data**. *Bioinformatics* 2002, **18 Suppl 1**:S136–144.

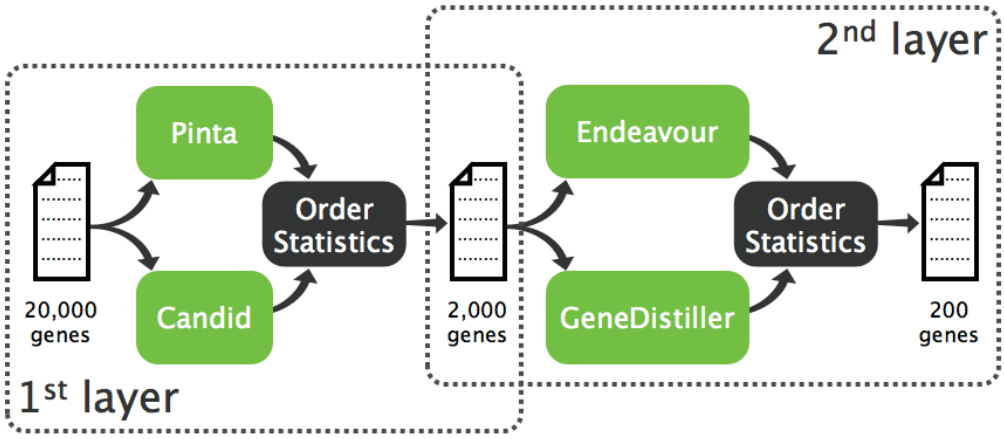


Figure 1 - Schematic representation of our two layer holistic approach

In the first layer, Pinta and Candid prioritize the whole human genome to define a set of 2,000 promising candidate genes. In the second layer, these genes are prioritized with Endeavour and GeneDistiller to identify the 200 most promising candidate genes.

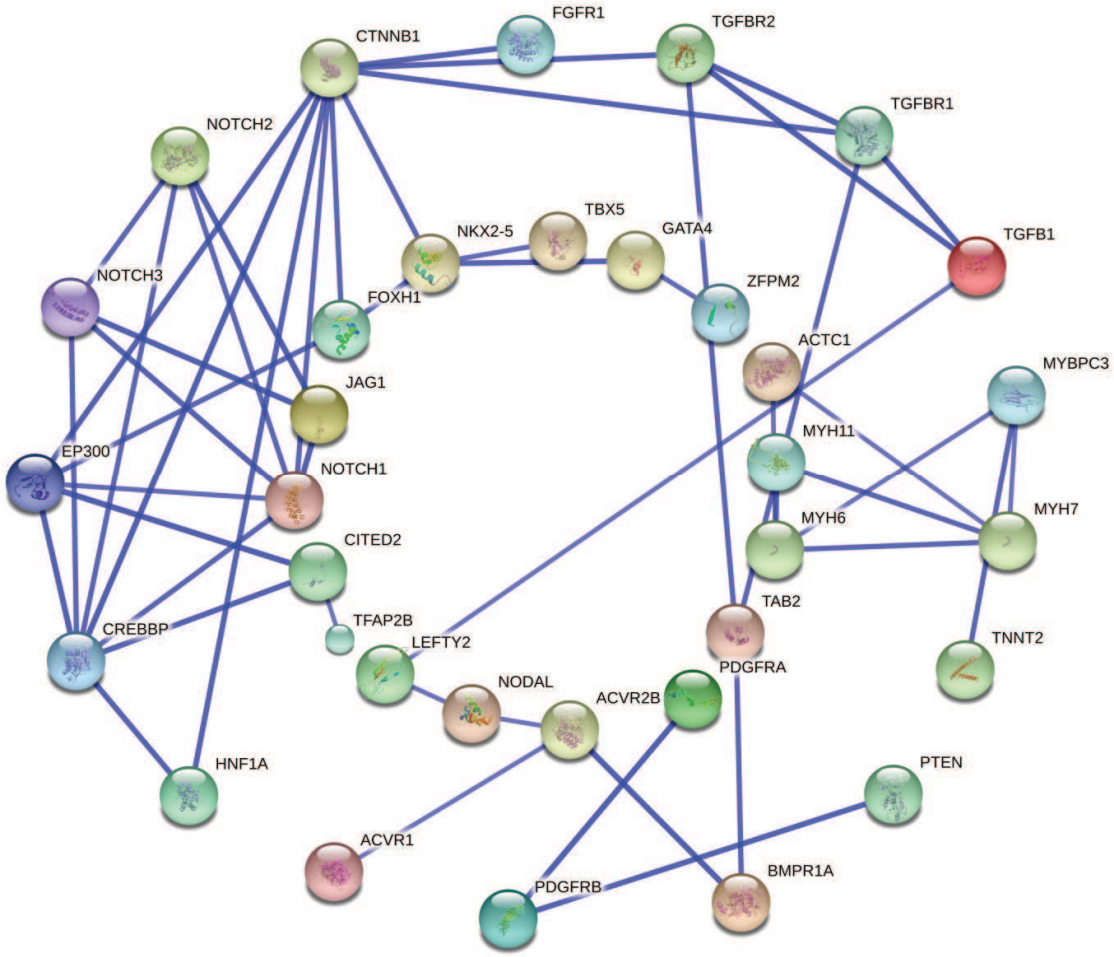


Figure 2 - Network view of the known CHD genes and the top predictions

Network view that combines the known congenital heart defect genes (inner circle) and the most promising candidates (outer circle). Genes without interactions have been removed. The displayed interactions are extracted from knowledge bases and experimental databases.

CHD gene set	Keyword Set
Gene name	**Term**
ACTC1	congenital heart defect
ACVR2B	heart
ALDH1A2	cardiac
ANKRD1	cardial
CFC1	cardiovascular
CITED2	valve
CRELD1	septum
ELN	atrioventricular
ZFPM2	atrium
FOXH1	atrial
GATA4	mitral
GATA6	aortic
GDF1	outflow
GJA1	systolic
LEFTY2	diastolic
MAP3K7IP2	endocardial
MYH11	splanchnic mesoderm
MYH6	anterior-posterior polarity
NKX2-5	left-right polarity
NKX2-6	neural crest
NODAL	looping
NOTCH1	coronary
PDGFRA	endocardial cushion
TBX20	hypertrophy
TFAP2B	cardiomyopathy
THRAP2	pericardium
ZIC3	epicardial
JAG1	epicardium
TBX5	myocardial
TBX1	cardiomyocyte
	septal defect
	tetralogy
	fallot
	perimembranous
	arteriosus
	cyanogenic
	cyanosis
	truncus
	trunk
	ductus
	eisenmenger
	dilated
	transposition
	hypoplastic left heart syndrome
	pulmonary insufficiency
	pulmonary stenosis

Table 1 – Training set and keywords for congenital heart defect

This table contains the training data used in our workflow *i.e.* the keyword set and seed gene set built for congenital heart defect

CDH gene set	CDH keyword set
Gene name	**Term**
STRA6	congenital diaphragmatic hernia
RBP1	bochdalek hernia
RBP2	left-side CDH
RBP3	isolated CDH
RBP4	congenital hernia of diaphragm
RBP5	eventration of the diaphragm
CRABP2	diaphragmatic hernia
RBP7	left-sided CDH
POLR2L	diaphragm eventration
AVP	diaphragm
ADH1C	diaphragmatic
ADH1B	diaphragmatic defect
ADH1A	posterolateral hernias
ADH4	pulmonary hypoplasia
ADH5	pulmonary hypertension
ADH7	retinoic acid pathway
ADH6	retinoic acid
RDH5	vitamin A
RDH8	retinoid
RDH10	retinol
RDH11	retinoid signalling pathway
RDH12	
RDH13	
RDH14	
DHRS9	
RDH16	
DHRS3	
LRAT	
ALDH1A1	
ALDH1A2	
ALDH1A3	
CRABP1	
CYP26A1	
CYP26B1	
RARA	
RARB	
RARG	
RXRA	
RXRB	
RXRG	
NR2F2	
ZFPM2	
GATA4	

Table 2 – Training set and keywords for congenital diaphragmatic hernia

This table contains the training data used in our workflow *i.e.* the keyword set and seed gene set built for congenital diaphragmatic hernia

Asthma gene set	Asthma keyword set
Gene name	**Term**
ORMDL3	asthma
ZPBP2	atopy
IL1RL1	airway hyperresponsiveness
IL18R1	FEV1
IL33	CP20 methacholine
SMAD3	allergic asthma
CX3CR1	airway inflammation
ALOX15	airway remodeling
PLAU	
FLG	
CD14	
PTPRE	
IL1R2	
BACE1	
IL1R1	
SPI1	
MMP2	
CAT	
NQO1	
C5	
NPSR1	
VDR	
IL13	
IL4	
IL4R	
IFNG	
ADAM33	
IL10	

Table 3 – Training set and keywords for asthma

This table contains the training data used in our workflow *i.e.* the keyword set and seed gene set built for asthma

Rank	Gene	Band	Disease evidence
1	BMPR1A	10q22	10q22-q23 deletion syndrome
2	MYH7	14q12	Atrial septal defect, Ebstein anomaly, cardiomyopathy
3	MYBPC3	11p11	Cardiomyopathy, left ventricular noncompaction
4	SCN5A	3p21	Cardiomyopathy, long QT syndrome
5	LMNA	1q22	Cardiomyopathy, left ventricular noncompaction
6	TGFBR2	3p22	Loeys-Dietz syndrome, Marfan syndrome
7	TGFBR1	9q22	Loeys-Dietz syndrome, Marfan syndrome
8	NOTCH2	1p12	Alagille syndrome
9	FGFR1	8p12	Kallman syndrome
10	NOTCH3	19p13	*Animal model: cardiac fibrosis*
11	TGFB1	19q13	Marfan syndrome
12	FBN1	15q21	Marfan syndrome, mitral valve prolapse
13	ACVR1	2q24	Atrial septal defects
14	TNNT2	1q32	Cardiomyopathy, septal defects
15	CREBBP	16p13	Rubinstein-Taybi syndrome, pulmonary valvar stenosis
16	CASQ2	1p13	Ventricular tachycardia
17	PDGFRB	5q33	*Animal model: neovascularization*
18	CTNNB1	3p21	*Animal model: cardiac development*
19	PTEN	10q23	*Animal model: left ventricular hypertrophy*
20	HAND1	5q33	Septal defects
21	COL1A1	17q21	*Animal model: heart valve disease*
22	HNF1A	12q24	*Animal model: atrioventricular septal defects*
23	EP300	22q13	Rubinstein-Taybi syndrome
24	LAMC1	1q31	*Animal model: cardiomyopathy, heart valve disease*
25	AKT1	14q32	*Animal model: coronary heart disease*

Table 4 – Most promising candidate genes for congenital heart defects

The most promising candidate genes for CHD are presented with functional evidence when applicable.

Loci	#genes in region	Top gene	Genome-wide rank
Chr 1 (170596692, 192187380)	167	ABL2	336
Chr 2 (230278105, 242923099)	175	HDAC4	394
Chr 4 (terminal, 2064607)	56	FGFR3	16
Chr 4 (139442723, 156051983)	127	EDNRA	125
Chr 4 (22766528, 36328463)	38	PPARGC1A	313
Chr 5 (terminal, 18712406)	112	TRIO	210
Chr 6 (terminal, 7193095)	74	FOXC1	283
Chr 6 (155239588, 170899992)	110	RPS6KA2	378
Chr 9 (terminal, 9402318)	79	JAK2	245
Chr 11 (114237274, 134452384)	276	MLL	165
Chr 14 (87952783, 104622155)	307	AKT1	72
Chr 15 (70485883, 73852471)	70	NEO1	168
Chr 22 (14506719, 24620247)	292	MAPK1	213
Chr X (120476201, 130520358)	51	XIAP	892

Table 5 - Analysis of CHD associated regions

First, a list of 14 known CHD loci is compiled. For each locus, the most promising candidate gene and its genome-wide ranking are indicated. Complete results are provided in Additional File 3.

Position	Gene	Association to asthma
1	*TGFB1*	*Associated with asthma in several studies* [94, 95]
2	*IL1B*	*Weakly associated with asthma in Canadian / Australian populations* [96]
3	*PTPRC*	Splicing of *PTPRC* is regulated by *SFRS8*, a gene associated with asthma [114].
4	*TGFBR2*	*Several polymorphisms associated with atopy and asthma* [98]
5	*TLR4*	*Associated with asthma in Swedish children* [99]
6	MYD88	Its over expression causes an increase in the level of transcription from interleukin-8 promoter [115]
7	*IL6*	*Variants modify the bronchodilator drug responsiveness in asthma* [100]
8	SMAD2	A key element on the TGFβ1 signaling pathway in human bronchial epithelial cells, that is altered in asthmatic bronchial epithelial cells [116]
9	IL1RAP	A transmembrane protein required for interleukine-1 and linked to IL-33 [117] and IRAK [118]
10	*TLR2*	*Determinant of susceptibility to asthma in children of European farmers* [101]
11	NFKB1	Linked to IL-8 [119]
12	CSF2RA	No association known to asthma
13	INPP5D	Associated with TLR4 [120], IL-8 [121] and TNFa [122]
14	CD4	$CD4^+$ T cells producing Th2 cytokines play a prominent role in the lungs of asthmatic subjects [123]
15	IL10RA	No association known to asthma
16	RELA	Indirectly linked to CCL3, IL23A, TNF and IL1B [124]
17	*CCL2*	*Associated with asthma, and asthma severity* [102, 103]
18	FAS	Related to autoimmune disease [125] and present in its soluble form (sFas) in allergic asthmatic patients [126]
19	IKZF1	Thought to play an important role in CD4 and CD8 lineage [127]
20	CREBBP	Linked to IL-4 [128]
21	IL2RG	No association known to asthma
22	ITGB2	Involved in induction and adhesion of eosinophils [129]
23	*IL12B*	*Associated with severity of atopic and non-atopic asthma* [104, 105]
24	*C3*	*Variants affect susceptibility to bronchial asthma* [106]
25	TNFRSF1A	No association known to asthma

Table 6 – Top 25 genes for asthma

A list of the top ranked 25 genes in asthma and their relation with the disease, if known, is presented.

Chapter 5

Summary

In this chapter, we present a work on gene prioritization where we propose a preprocessing method to enhance the ranking capacity of training set based gene prioritization tools.

A group of gene prioritization tools use training genes to lead the prioritization process. These genes, usually identified as linked to the genetic condition under investigation, provide the tool with a profile which will be later used to classify the candidate genes. In these tools, including Endeavour[4] and Genedistiller[3], the selection of the training genes entirely depends on the expertise of the user.

Our intention with this work is to relieve the user from the complete responsibility when selecting genes for the training set. We intend to do so using cluster analysis to find groups of similar genes in the training set.

Due to the categorical nature of the databases used in the analysis, we have selected a transactional based clustering algorithm: CLOPE. This algorithm uses a global measure which depends on the histograms of the clusters and it has been successfully applied to market basket type problems.

We have applied cluster analysis to 27 expert selected training sets and to the only training set based gene prioritization tools that allow massive experiments, Endeavour and Genedistiller.

We have compared the validation results of the expert selected training sets and the ones returned by the cluster analysis and we have seen a general increase in the quality of the ranking. To discard the reduction of size as a cause of the improvement, we have compared randomly selected training sets with the clustered ones and the results also show a general improvement after the clustering.

This work has been submitted to BMC Bioinformatics On May 2012

Personal contribution

The Ph.D. candidate has set up the initial idea, devised the strategy, implemented the cluster analysis, analyzed the results and finally written the manuscript.

A clustering based preprocessing method for gene prioritization

Francisco Bonachela-Capdevila[1], Léon-Charles Tranchevent [2], Yves Moreau[2], Patrick De Causmaecker[1§]

[1]CODeS Group, ITEC-IBBT-KULEUVEN, Katholieke Universiteit Leuven campus Kortrijk, Kortrijk, Belgium

[2]ESAT-SCD / IBBT-K.U.Leuven Future Health Department, Katholieke Universiteit Leuven, Leuven, Belgium

[§]Corresponding author

Email addresses:

FBC: Francisco.BonachelaCapdevila@kuleuven-kortrijk.be

LCT: Leon-Charles.Tranchevent@esat.kuleuven.be

YM: Yves.Moreau@esat.kuleuven.be

PDC: Patrick.DeCausmaecker@kuleuven-kortrijk.be

Abstract

Background

Unravelling the causes of polygenic diseases is an essential aspiration of human genetics. Bioinformaticians have defined the candidate gene prioritization problem and several tools in order to tackle this problem have been developed in the last years. A group of these tools, Endeavour and GeneDistiller included, rely on a so-called training set of genes, based on the expertise of the user to rank candidate genes. However, when complex diseases or general terms including different subtype syndromes are validated, the results are not as good as the ones obtained with more simple, yet polygenic, diseases.

Results

To address this challenge we have applied clustering as a pre-processing step in gene prioritization to obtain more homogeneous training sets and, therefore, more accurate rankings. Due to the nature of the biological data during the cluster analysis, we have selected CLOPE, a clustering algorithm designed to cope with transactional data.

Conclusions

Leave-one-out cross-validation experiments show that the use of clustering as a pre-processing step in training set based gene prioritization leads, in general, to smaller and more homogeneous training sets and to a more accurate final ranking.

Background

Very common and life taking diseases in our society, such as cardiovascular disorders, diabetes, schizophrenia and numerous forms of cancer, among other conditions, are controlled by both environmental and genetic factors. Identifying the molecular basis of these complex conditions is essential since it is a critical step towards the design of more reliable diagnostic tests and the development of more effective treatments.

Conventional strategies like linkage analysis and positional cloning have been successfully applied in the last decades to unravel the genetic basis of monogenic diseases and to identify the underlying causative genes (Altshuler et al., 2008). More recently, Genome Wide Association Studies (GWAS) have been used to identify associations between chromosomal loci and complex human diseases (Cantor et al., 2010). However, due to the complex nature of the polygenic disorders, these techniques often fail to identify the exact disease causing gene (The Wellcome Trust Case Control Consortium, 2007). Indeed for complex diseases, the genetic mechanism that triggers the abnormal phenotype is shared by several genes that are acting in conjunction, meaning that the individual effect of a single gene is weaker than for monogenic conditions. Therefore, traditional techniques tend to identify the chromosomal regions that harbour the genes rather than the genes themselves. These regions can be very large, ranging from kilobases for genetic association studies up to megabases for genetic linkages, and often contain hundreds of genes (Hardy et al., 2009). Usually only one or a few of these genes are of primary interest, and identifying these novel disease associated genes can be an expensive and time-consuming task if all the candidate genes must be individually experimentally validated.

Consequently, bioinformaticians have plunged into this problem through the definition of the candidate gene prioritization problem and the development of several prioritization methods (reviewed in Tranchevent et al., 2011). Selecting the most promising candidate genes among a large pool of candidate genes is nowadays facilitated by computational tools that take advantage of both the fast development of informatics and the exponential growth of the biological databases they are using. The vast majority of these tools select promising candidates using the guilt-by-association concept: the best candidate genes are the ones that are similar to the genes already known to be involved in the disease under study. Thus, a key step for an efficient prioritization is the identification of these genes already associated to the disease under study, that together form a training set. However, results are more accurate if the training set is homogeneous and describes very precisely one single biological process. At the contrary, prioritization is less effective if the training set is heterogeneous or contains outliers. In the latter case, the genes do not share enough relevant features to build an accurate model. For example, when working on complex disorders such as congenital heart defects, more accurate

results are obtained when prioritizing with seven training sets that cover the different biological processes associated to congenital heart defects (Thienpont et al., 2010).

In the present study, we test the hypothesis that the performance of training set based gene prioritization methods is influenced by the homogeneity of the training sets, and that therefore clustering based pre-processing can enhance their performance. To this end, we propose a clustering based pre-processing strategy that identifies one or several homogeneous training sets from a single potentially heterogeneous training set. The effectiveness of the proposed strategy is then assessed through a benchmark analysis on known data.

Results

Clustering based pre-processing

In this paper, we develop a clustering based strategy that is applied prior to prioritization to remove outliers and to obtain clusters that are more homogeneous than the original gene set (see Figure 1). We perform clustering on 29 disease gene sets using CLOPE. Clusters with less than 3 elements are discarded, and therefore, only 15 of the 29 diseases return valid clusters. In addition, for 5 of these 15 diseases, more than one cluster is produced. Table 1 presents a summary of the clustering results for the 15 diseases that returned valid clusters (see also Table B.2 in appendix B for the complete results). The disease sets contain 27 genes on average, and clusters are on average reduced to 11 genes, which represents a reduction of 60%. We can observe that 92,2% of the discarded clusters contain a single gene, the remaining 7,8% contain two genes.

Results of the benchmark

We assess the relevance of our clustering strategy through a cross-validation scheme on known data using two prioritization methods termed Endeavour and GeneDistiller. More precisely, we perform 25 replicates of leave-one-out cross-validation for both the original disease gene sets and the clusters issued from them (see Figure 1). The classification error values (AACs) for this benchmark and the average of their medians are presented in Figure 2 (see also Tables B.3 and B.4 in appendix B for the numeric values and Tables B.5 and B.6 also in appendix B for the numeric values for diseases with more than one cluster). We can observe that, in general, the AACs values of the clusters are smaller than those of the original disease gene sets. This is observed for 12 and 10 of the 15 diseases for Endeavour and Genedistiller.

Control experiment

We then assess whether the observed increase in performance is a consequence of the cluster homogeneity or of the size reduction. To this end, we also cross-validate

randomly generated clusters and compare their AACs to the AACs obtained with the real clusters. The Figure 2 contains the AACs for the 15 diseases for both prioritization methods (see also Table B.7 in appendix B for the complete results). We observe better results in general for the real clusters than for the control clusters in 14 and 10 of the 15 diseases respectively for Endeavour and GeneDistiller.

Discussion

This paper describes the application of a clustering method as a pre-processing step to enhance human candidate gene prioritization. Several disease specific gene sets are refined using a clustering strategy. A benchmark analysis shows that our unsupervised classification can generate smaller and more homogeneous gene sets that lead in general to better performance.

The original benchmark on 29 genetic disorders has revealed that the cross-validation performance is significantly lower than the average for a few diseases (*e.g.*, Alzheimer's disease, amyotrophic lateral sclerosis). A reasonable explanation is the possible heterogeneity of the corresponding training sets mainly because disease names are sometimes broad terms that include distinct disease subtypes. In addition, the worst performance is sometimes associated to diseases with large training sets, that gather together genes from multiple pathways. For example, the leukemia gene set contains 112 genes, however several types of leukemia can be distinguished: Acute Lymphoblastic Leukemia (ALL), Chronic Lymphocytic Leukemia (CLL), Acute Myelogeneous Leukemia (AML), and Chronic Myelogeneous Leukemia (CML) among others. Another example is neuropathy that comprises a broad range of syndromes that affect nerves and/or nerve cells possibly through the disturbance of distinct pathways (Reilly, 2009).

We use CLOPE for our clustering approach because it is well suited for our genomic data. In addition, its performance in terms of computing time makes it suitable for our benchmark. The clustering is based on functional annotations that describe accurately the current knowledge we have about the genes and their functions. Our clustering method only uses the annotations that are significantly over-represented within the disease gene set as compared to the whole genome. The resulting clusters are therefore only based on annotations that are relevant to the disease under study. Our results show that our clustering approach also performs outlier detection. Indeed, 40% of the genes present in the original disease gene sets are considered as outliers and therefore removed prior to further analysis. We notice that 92,2% of these outliers are discarded because the corresponding clusters only contain a single element. These genes are discarded because they do not share annotations with the other genes and keeping them would hinder the homogeneity of the cluster. The remaining 7,8% outliers are issued from clusters of size two. These clusters are too small to undergo a leave-one-out cross-validation since that would leave only one gene for training when the other is left out. It does not mean however

that these clusters are irrelevant for real predictive studies, for which the two genes can be used for training. We benchmark our strategy using the 15 diseases for which valid clusters are obtained. An explanation for such reduction lies in the nature of the clustering algorithm and the heterogeneity of the databases. The algorithm strongly rewards genes that are very similar across all databases (high histograms) but thus penalizes genes for which no consensus can be found across all databases. Missing values are also an important cause for this stringency and since not all five databases have the same number of records, genes with missing data are strongly penalized. Another limitation is related to the clustering procedure. Most clustering algorithms, including CLOPE, place one gene in a single cluster. With this simple assumption, it is impossible to model accurately the fact that one gene can participate in several pathways and therefore should ideally be part of several clusters. Future approaches to this problem should include the use of clustering strategies capable of producing overlapping clusters.

A closer look at the clustering results for myopathy reveals interesting insights.

The two myopathy clusters correspond to two different situations. Cluster 2 contains the following genes: *COL6A3, COL6A2, COL6A1*, and *COL9A3*. These genes are all members of the collagen superfamily and the first three of them are associated to a particular type of myopathy: Bethlem myopathy. At contrary, cluster 1 contains *TNNC* and *ACTC1*, both linked to dilated cardiomyopathy, and *MYH2* and *MYL2*, two myosin genes. In the case of anemia, the cluster consists of five genes related to Fanconi anemia (*FANCD2, FANCE, FANCF, FANCC*) and in cataract, three crystalline proteins are clustered (*CRYGC, GRYGD* and *CRYBA1*). This illustrates that the clusters we produce make sense biologically, however this does not hold for all diseases.

To benchmark our approach, we have used two gene prioritization methods: our tool Endeavour, and GeneDistiller for which an API is available. These methods both rely on a guilt-by-association model to compute similarities between known genes and candidate genes. Therefore, the databases used to cluster the gene sets have not been used for prioritization to avoid biased results. Our benchmark indicates that the performance is higher for the clusters than for the original gene sets, and this for both prioritization methods. For some diseases, however, the classification error is larger after clustering. This is mostly the case for diseases with small training sets (*e.g.*, spinocerebellar ataxia, spastic paraplegia, Ehlers-Danlos syndrome) that are already more homogeneous than the larger gene sets of the more complex diseases (*e.g.*, leukemia, anemia). Another explanation is that the databases used for clustering and the ones used for prioritization are different, and it is possible that what is identified as a cluster using a first set of databases does not represent a homogeneous gene set when other databases are considered. One option to circumvent that problem is to include more databases in the clustering process. Furthermore, a comparison between our valid clusters and randomly generated sets

of genes of the same size indicates that the performance improvement is not due to size reduction. For Endeavour, on average, the performance for randomly generated clusters (8,27%) is even worse than the performance for complete disease set (6,24%), indicating that size reduction does not lead to better performance in general.

One limitation of our approach is the correlation between the annotation terms, that is not corrected. As a result, the clustering process is potentially driven by correlated features, which might bias the approach. In the future, we want to extend the analysis by starting with a set of uncorrelated annotations.

Conclusions

We have developed a clustering based pre-processing method for human candidate gene prioritization. A benchmark analysis has revealed that the resulting clusters lead to better performance because the prioritization methods take advantage of their intrinsic homogeneity. In particular, for complex disorders with several subtypes or distinct associated phenotypes, the use of several clusters automatically derived from a large disease gene set, allows for a more precise modelling of the disease, and, presumably, more accurate predictions.

We have used five different databases to cluster the genes, we consider our priority to extend this set of databases to cover other gene characteristics. With more databases added, we expect a clearer link between different syndromes of a disease and clusters.

Methods

Clustering data

Biological databases can be classified into two categories. On the first hand, knowledge bases contain structured data that describe our current knowledge in genetics and molecular biology. On the other hand, huge data repositories, such as the gene expression omnibus (GEO), contain raw data that can be used to discover novel patterns in the data. Traditionally, the combination of these two types of data provides a good balance between reliability and novelty. For our study, knowledge bases such as Gene Ontology and Kegg are ideal since we are looking for functionally homogenous clusters. These biological databases often use different data representation (*e.g.*, vector based, network based) with different properties (*e.g.*, binary values, continuous values, various levels of sparseness). It has been shown that cluster analysis is extremely dependent on the type of attributes of the elements to be clustered and mixing different types of data normally lead to less accurate results (Ahmad et al., 2007). Therefore, for the present study, we have chosen to focus on the most prevalent type of data sources in order to use a single clustering algorithm. To this end, we have selected the annotations based data

sources underlying Endeavour, whose information is stored as binary vectors: Gene Ontology (The Gene Ontology Consortium 2000), KEGG (Kanehisa et al., 2000), InterPro (Hunter et al., 2009), EnsemblEST (Hubbard et al., 2009) and SwissProt (The Uniprot Consortium, 2007).

When a gene set is considered, not all annotations should be treated equally since some annotation attributes are frequently found genome-wide and others are rare. Also, as the annotation attributes are organized in tree structures, there is redundancy between any parent attribute and its children. Last, genes with multiple functions are annotated with multiples terms, possibly including terms that are not related to the disease. To circumvent these problems, only the significantly over-represented attributes, retrieved using a method described in Aerts et al., 2006, are used. Briefly, the over-representation of each term within the training set as compared to the genome is assessed independently using the binomial distribution; and the Bonferroni step-down method is used to correct for multiple testing. Every gene is then represented by a binary profile that only contains the most relevant attributes. The clustering task is then performed on these vectors.

Clustering algorithm

To account for the categorical nature, the high dimensionality, and the sparseness of our data, we have selected a clustering algorithm that focuses on categorical attributes and that is not based on the concept of distance, which could not be used due to the high number of dimensions (curse of dimensionality, Korn et al., 2001). We have chosen CLOPE, a clustering algorithm designed to cope with transactional data of high dimensionality (Yang et al., 2002). This clustering algorithm works by assigning every element to a cluster maximizing the height-to-width ratio of the histogram of every cluster. Applied to our situation, this means that the more attributes that two genes share and the less non-shared attributes that every particular gene has, the more similar the two genes will be. Therefore, clusters containing genes that share as many attributes as possible (height maximization) and which have as few attributes as possible which are poorly shared (width minimization) are preferred. The profit function (1) which leads the clustering process depends on the area of the histogram (S), or, what is the same, the height (H) by the width (W) of every histogram as well as on the size (|C|) and amount of clusters (k):

$$Profit_r(\boldsymbol{C}) = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} * |C_i|}{\sum_{i=1}^{k} |C_i|}$$

CLOPE only has a single parameter: the repulsion coefficient *r*. This coefficient helps the algorithm to adapt to different degrees of sparseness. The acceptance criterion of CLOPE for two genes to be in the same cluster also depends on the distribution of the similarities between all gene pairs. This strategy of building the clusters based on the shape of the histograms of every cluster instead of a concept of distance turns CLOPE into a suitable algorithm to deal with our data. To select the optimal repulsion during the clustering process, and eventually, the number of clusters, we have run CLOPE with every possible repulsion value between 0 and the maximum value (that corresponds to the repulsion value that splits the original cluster into *n* single-gene clusters, being *n* the total number of genes), with increments of 0.1. For every disease, the value that returned the highest profit function was selected for our benchmark. Table B.1 in appendix B contains the repulsion values for the 15 diseases. Clusters containing a single element are regarded as outliers and discarded. Clusters of two elements are also regarded as outliers even if they would make sense biologically. This approximation is necessary since Endeavour and GeneDistiller do not work optimally with too small training sets and because our benchmark is based on a leave-one-out scheme.

Gene prioritization methods

Gene prioritization is the process of identifying the most promising candidate genes from a large set of candidate genes with respect to a disease of interest. Many computational methods have been developed to tackle that problem. Different algorithms can be utilized, various combinations of biological databases can be defined, and thus dozens of different methods have been developed. For this study , we focus on the methods that start from a set of already known disease genes (also termed training genes) and then score the candidate genes based on their similarities to these already known genes. We benchmark our clustering approach using two prioritization methods Endeavour and GeneDistiller.

We have previously developed Endeavour, a tool that prioritizes candidate genes based on their similarities to a set of known disease genes (Aerts et al., 2006). The algorithm uses different data sources, including functional annotations, large expression datasets from Gene Atlas, literature data through text mining, regulatory information, and interaction networks. Endeavour uses a three step algorithm. The first step is the training in which information about the known disease genes is extracted to build models of the disease under study (one model per data source). In the second step, the candidate genes are scored for each model and ranked accordingly. The third step is the fusion, via the order statistics, of these complementary and sometimes contradictory rankings into one global ranking that determines the most promising candidate genes. A benchmark study through leave-one-out cross-validation on known disease and pathways genes showed that, when left out, the correct gene was ranked, on average, 10th out of 100 genes. We further experimentally validated Endeavour through the identification of 12 novel atonal

mediated neural development genes in Drosophila melanogaster (Aerts et al., 2009), and through the identification of a novel gene involved in heart development TAB2 (Thienpont et al., 2010).

GeneDistiller also relies on heterogeneous sources of information, including functional annotations, protein interactions, genetic markers, protein domains, families and paralogs, and phenotype information (Seelow et al., 2008). It uses various strategies to guide the process of gene prioritization: either through the use of keywords describing the disease, or through expression data or, similarly to Endeavour, through the definition of a set of known disease genes. In addition, an API is available for long or time-consuming queries, which is required for efficient benchmarking. Of interest, GeneDistiller has contributed to the discovery of a novel gene for infantile cerebral and cerebellar atrophy *MED17* (Kaufmann et al., 2010).

The biological data sources used for the clustering process have not been used for the prioritization procedure to avoid redundancy. The remaining databases contain protein-protein interactions (BIND (Bader et al., 2001), BioGrid (Stark, 2006), HPRD (Peri et al., 2003), InNetDb (Xia et al., 2006), IntAct (Kerrien et al., 2007), MINT (Chatr-aryamontri et al., 2007), STRING (Von Mering et al., 2007), UniHI (Chaurasia et al., 2006)), expression data (Su et al., 2002; Son et al., 2005), mitochondrial protein specific data (Maestro (Calvo et al., 2006), Mitropred (Guda et al., 2004)), global disease probability scores (Lopez-Bigas et al., 2004; Adie et al., 2005), genetic markers (dbSNP (Sherry et al., 2001), UniSTS (Wheeler et al., 2008)), literature data (Glenisson et al., 2004) and sequence similarities (Ye et al., 2006).

Benchmark procedure

For this study, we have used an already described benchmark dataset (Aerts et al., 2006). This dataset consists of 620 genes distributed in 29 distinct diseases. In a first step, each of the 29 disease gene sets is clustered individually using CLOPE in a WEKA environment (Hall et al., 2009).

We estimate the efficiency of our clustering approach with a leave-one-out cross-validation (see Figure 1). More precisely, for each disease, we cross-validate the original disease gene set and the clusters derived from it, and compare the results. In a leave-one-out setup, all disease genes except one (termed the defector gene) are used for training (*i.e.*, positive genes). The candidate set then contains 99 randomly chosen genes plus the defector gene (*i.e.*, unlabelled genes). The position of the defector genes among the 100 candidate genes indicates how well the computational method is able to detect its association to the disease under study (see Figure 1). This procedure is repeated for all the disease genes so that each gene is, in turn, left-out. After these repetitions, we have as many rankings as there are genes in the set. We then compute a ROC curve by investigating these rankings with a varying threshold. For each threshold value, the sensitivity and specificity are

calculated, and a point is drawn in the ROC space. By varying the threshold, a complete Receiver Operating Characteristic (ROC) curve can be built. The Area Above the Curve (AAC) is used as an estimate of the classification error, therefore allowing the comparison between different validations. The AAC ranges from 0 (*i.e.*, perfect prioritization) to 1 (*i.e.*, perfectly inverse prioritization) with 0.5 representing the random expectation. This validation procedure is performed for all 29 diseases, and then repeated 25 times to get an estimate of the variance. For each disease, a candidate set consisting of 99 randomly chosen genes is built beforehand and used for both the disease and the corresponding clusters to be able to derive fair comparisons. Notice however that a different candidate set is used for each of the 25 repetitions.

In addition, and as a control, random clusters are generated and benchmarked following the above procedure. Random clusters are of the same sizes as the original clusters and are created from the same disease gene sets. For each real cluster, 10 random clusters are built and cross-validated 25 times using the previously defined candidate sets.

Authors' contributions

The work presented here has been carried out in collaboration between all authors. FBC initiated the approach, designed the experiments, performed the cluster analysis, the validation experiments using GeneDistiller and wrote the manuscript. LCT co-designed the experiments, performed the validation tests with Endeavour and wrote the manuscript. YM co-designed the experiments and discussed analysis. PDC initiated the approach, co-designed the experiments, discussed analysis and coordinated the project. All authors read and approved the final manuscript.

Author's information

Francisco Bonachela-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.

Léon-Charles Tranchevent is a post-doctoral fellow 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.

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.

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.

Acknowledgements

We thank Prof. Hans Deckmyn, head of the Interdisciplinair Research Centrum, K.U. Leuven and Prof. Dr. Peter Vandenberghe, from U.Z. Leuven for their help in unraveling the biological meaning of our clusters. We deeply appreciate the assistance of Prof. Dr. med. Markus Schülke and Dominik Seelow with GeneDistiller's API. This work was supported by BIOPTRAIN, K.U.Leuven [K.U. Leuven CIF/07/02], Research Council KUL (ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys, KUL PFV/10/016 SymBioSys, START 1), FWO (G.0318.05, G.0553.06, G.0302.07, ICCoS, ANMMM, MLDM, G.0733.09, G.082409), IWT (Silicos, SBO-BioFrame, SBO-MoKa, TBM-IOTA3), FOD:Cancer plans, IBBT, Belgian Federal Science Policy Office (IUAP P6/25), EU-RTD (ERNSI, FP7-HEALTH CheartED).

References

1. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization**. *BMC Bioinformatics* 2005, **6**:55.

2. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y: **Gene prioritization through genomic data fusion**. *Nat Biotech* 2006, **24**:537-544.

3. Aerts S, Vilain S, Hu S, Tranchevent L-C, Barriot R, Yan J, Moreau Y, Hassan BA, Quan X-J: **Integrating Computational Biology and Forward Genetics in Drosophila**. *PLoS Genet* 2009, **5**:e1000351.

4. Ahmad A, Dey L: **A k-mean clustering algorithm for mixed numeric and categorical data**. *Data & Knowledge Engineering* 2007, **63**:503-527.

5. Altshuler D, Daly MJ, Lander ES: **Genetic mapping in human disease**. *Science* 2008, **322**:881-888.

6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A,

Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene Ontology: tool for the unification of biology**. *Nat Genet* 2000, **25**:25-29.

7. Audo I, Bujakowska K, Mohand-Saïd S, Tronche S, Lancelot M-E, Antonio A, Germain A, Lonjou C, Carpentier W, Sahel J-A, Bhattacharya S, Zeitz C: **A novel DFNB31 mutation associated with Usher type 2 syndrome showing variable degrees of auditory loss in a consanguineous Portuguese family.** *Mol Vis* , **17**:1598-1606.

8. Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CWV: **BIND—The Biomolecular Interaction Network Database**. *Nucleic Acids Research* 2001, **29**:242 -245.

9. Calvo S, Jain M, Xie X, Sheth SA, Chang B, Goldberger OA, Spinazzola A, Zeviani M, Carr SA, Mootha VK: **Systematic identification of human mitochondrial disease genes through integrative genomics**. *Nat. Genet* 2006, **38**:576-582.

10. Cantor RM, Lange K, Sinsheimer JS: **Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application**. *The American Journal of Human Genetics* 2010, **86**:6-22.

11. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database**. *Nucleic Acids Res* 2007, **35**:D572-574.

12. Chaurasia G, Iqbal Y, Hänig C, Herzel H, Wanker EE, Futschik ME: **UniHI: an entry gate to the human protein interactome**. *Nucleic Acids Res* 2007, **35**:D590-D594.

13. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization**. *Nucleic Acids Research* 2009, **37**:W305-W311.

14. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS: **PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites**. *Nucleic Acids Research* 2008, **36**:W399-W405.

15. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucl. Acids Res.* 2002, **30**:207-210.

16. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA: **Analysis of protein sequence and interaction data for candidate disease gene prediction**. *Nucleic Acids Res* 2006, **34**:e130.

17. Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B: **TXTGate: profiling gene groups with text-based information**. *Genome Biol* 2004, **5**:R43.

18. Guda C, Guda P, Fahy E, Subramaniam S: **MITOPRED: a web server for the prediction of mitochondrial proteins**. *Nucleic Acids Res* 2004, **32**:W372-W374.

19. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA data mining software: an update**. *SIGKDD Explor. Newsl.* 2009, **11**:10–18.

20. Han J, Kamber M: *Data Mining: Concepts and Techniques*. 1st edition. Morgan Kaufmann; 2000.

21. Hardy J, Singleton A: **Genomewide association studies and human disease**. *N. Engl. J. Med* 2009, **360**:1759-1768.

22. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM: **Using literature-based discovery to identify disease candidate genes**. *Int J Med Inform* 2005, **74**:289-298.

23. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P: **Ensembl 2009**. *Nucl. Acids Res.* 2009, **37**:D690-697.

24. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C: **InterPro: the integrative protein signature database**. *Nucl. Acids Res.* 2008:gkn785.

25. Hutz JE, Kraja AT, McLeod, HL, Province MA: **CANDID: a flexible method for prioritizing candidate genes for complex human traits**. *Genet. Epidemiol.* 2008, **32**:779-790.

26. Jain AK, Murty MN, Flynn PJ: **Data clustering: a review**. *ACM Comput. Surv.* 1999, **31**:264–323.

27. Jain AK: **Data clustering: 50 years beyond K-means**. *Pattern Recogn. Lett.* 2010, **31**:651–666.

28. Jonathan KB, Goldstein J, Ramakrishnan R, Shaft U: **When Is "Nearest Neighbor" Meaningful?** *IN INT. CONF. ON DATABASE THEORY* 1999:217--235.

29. Ju T, Warren J, Eichele G, Thaller C, Chiu W, Carson J: **A geometric database for gene expression data**. In *Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. Aachen, Germany: Eurographics Association; 2003:166-176.

30. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**:27-30.

31. Kaufmann R, Straussberg R, Mandel H, Fattal-Valevski A, Ben-Zeev B, Naamati A, Shaag A, Zenvirt S, Konen O, Mimouni-Bloch A, Dobyns WB, Edvardson S, Pines O, Elpeleg O: **Infantile cerebral and cerebellar atrophy is associated with a mutation in the MED17 subunit of the transcription preinitiation mediator complex**. *Am. J. Hum. Genet* 2010, **87**:667-670.

32. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H: **IntAct--open source resource for molecular interaction data**. *Nucleic Acids Res* 2007, **35**:D561-565.

33. Kohler S, Bauer S, Horn D, Robinson P: **Walking the Interactome for Prioritization of Candidate Disease Genes**. *The American Journal of Human Genetics* 2008, **82**:949-958.

34. Korn F, Pagel B-U, Faloutsos C: **On the "dimensionality curse" and the "self-similarity blessing."***Knowledge and Data Engineering, IEEE Transactions on* 2001, **13**:96-111.

35. López-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be involved in human genetic disease**. *Nucleic Acids Research* 2004, **32**:3108 - 3114.

36. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining**. *Nat. Genet* 2002, **31**:316-319.

37. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M,

Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A: **Development of human protein reference database as an initial platform for approaching systems biology in humans**. *Genome Res* 2003, **13**:2363-2371.

38. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ: **Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions**. *PLoS Genet* 2009, **5**:e1000534.

39. Reilly MM: **Classification and diagnosis of the inherited neuropathies**. *Ann Indian Acad Neurol* 2009, **12**:80-88.

40. Seelow D, Schwarz JM, Schuelke M: **GeneDistiller—Distilling Candidate Genes from Linkage Intervals**. *PLoS ONE* 2008, **3**:e3874.

41. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**:308-311.

42. Son CG, Bilke S, Davis S, Greer BT, Wei JS, Whiteford CC, Chen Q-R, Cenacchi N, Khan J: **Database of mRNA gene expression profiles of multiple human organs**. *Genome Research* 2005, **15**:443 -450.

43. Stark C: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Research* 2006, **34**:D535-D539.

44. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes**. *Proc. Natl. Acad. Sci. U.S.A* 2002, **99**:4465-4470.

45. The UniProt Consortium: **The Universal Protein Resource (UniProt)**. *Nucl. Acids Res.* 2008, **36**:D190-195.

46. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**:661-678.

47. Thienpont B, Zhang L, Postma AV, Breckpot J, Tranchevent L-C, Van Loo P, Møllgård K, Tommerup N, Bache I, Tümer Z, van Engelen K, Menten B, Mortier G, Waggoner D, Gewillig M, Moreau Y, Devriendt K, Larsen LA: **Haploinsufficiency of TAB2 causes congenital heart defects in humans**. *Am. J. Hum. Genet* 2010, **86**:839-849.

48. Tranchevent L-C, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y: **A guide to web tools to prioritize candidate genes**. *Briefings in Bioinformatics* 2011, **12**:22 -32.

49. Van Vooren S, Thienpont B, Menten B, Speleman F, De Moor B, Vermeesch J, Moreau Y: **Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations**. *Nucleic Acids Res* 2007, **35**:2533-2543.

50. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P: **STRING 7--recent developments in the integration and prediction of protein interactions**. *Nucleic Acids Res* 2007, **35**:D358-362.

51. Wang K, Xu C, Liu B: **Clustering transactions using large items**. In *Proceedings of the eighth international conference on Information and knowledge management*. 1999:490.

52. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2006, **34**:D173-180.

53. Xia K, Dong D, Han J-DJ: **IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model**. *BMC Bioinformatics* 2006, **7**:508.

54. Xiong Q, Qiu Y, Gu W: **PGMapper: a web-based tool linking phenotype to genes**. *Bioinformatics* 2008, **24**:1011 -1013.

55. Yang Y, Guan X, You J: **CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data**. *IN: PROC OF KDD'02* 2002, **2002**:682--687.

56. Ye J, McGinnis S, Madden TL: **BLAST: improvements for better sequence analysis**. *Nucleic Acids Res* 2006, **34**:W6-W9.

57. Yu W, Wulf A, Liu T, Khoury M, Gwinn M: **Gene Prospector: An evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases**. *BMC Bioinformatics* 2008, **9**:528.

58. Zenteno JC, Buentello-Volante B, Ayala-Ramirez R, Villanueva-Mendoza C: **Homozygosity mapping identifies the Crumbs homologue 1 (Crb1) gene as responsible for a recessive syndrome of retinitis pigmentosa and nanophthalmos**. *Am. J. Med. Genet.* 2011, **155**:1001-1006.

Figures

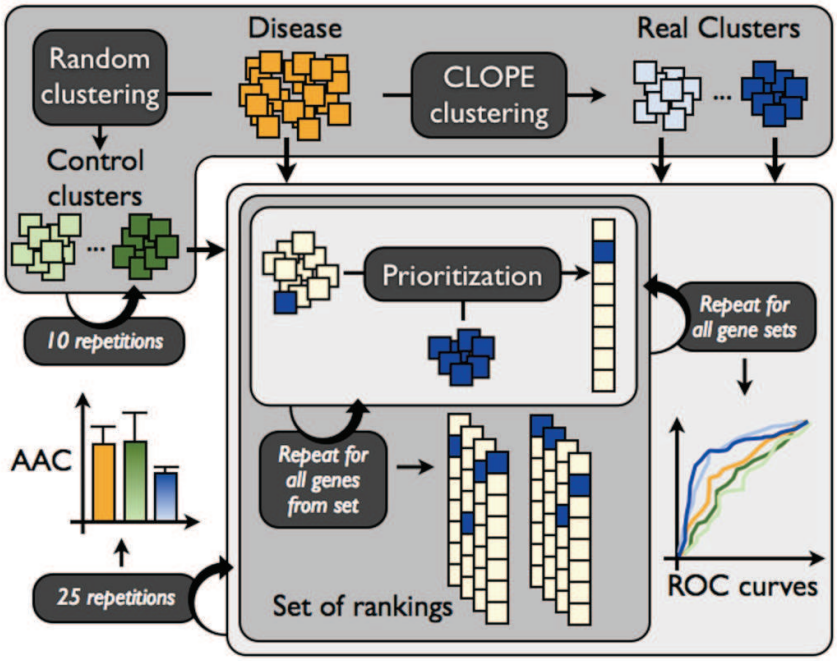


Figure 1 - Our validation workflow

Clustering is applied as a preprocessing step for each disease. In addition, control clusters are randomly generated. In a second step, a leave-one-out cross-validation procedure on know disease data is used to estimate the usefulness of the approach.

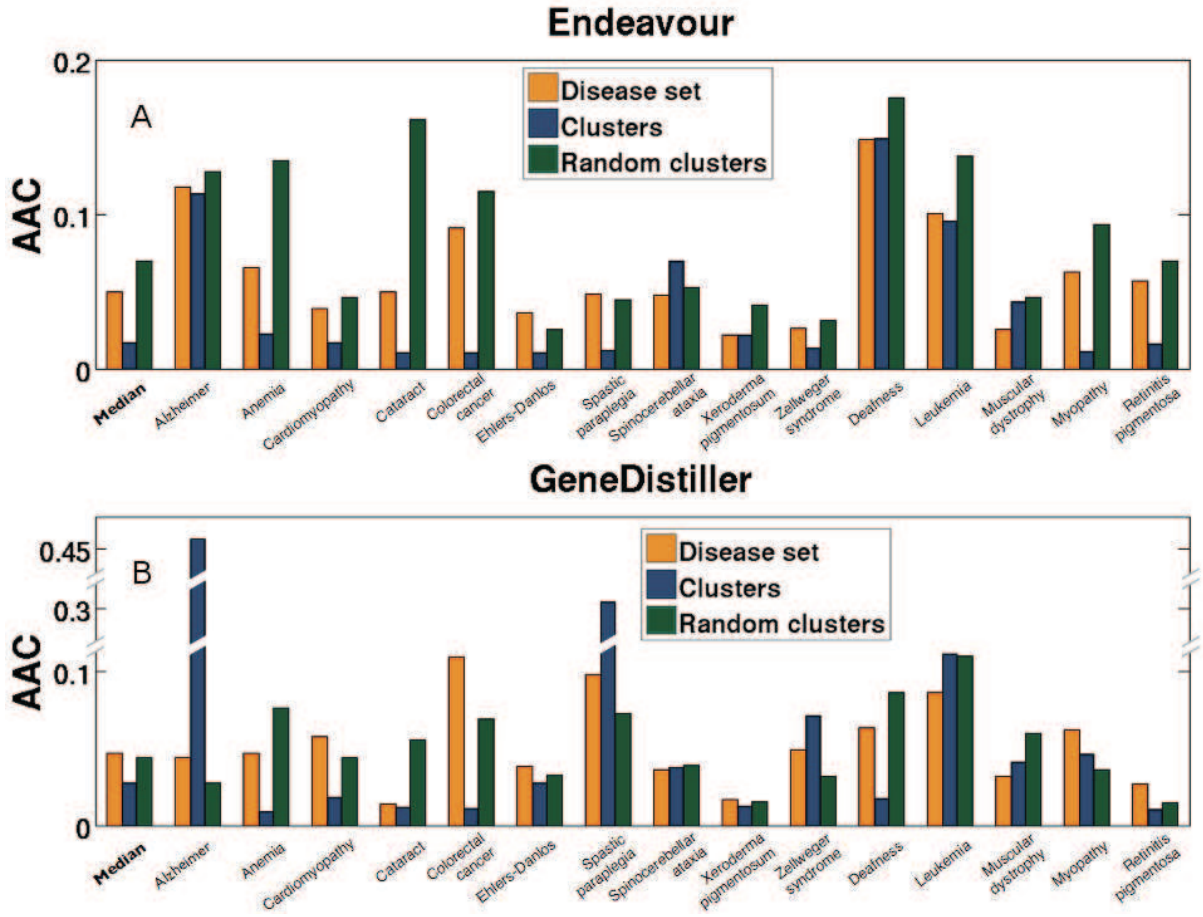


Figure 2 - Benchmark results

Performance of our disease based benchmark for Endeavour (A) and GeneDistiller (B). For each disease, the AACs are displayed for the original disease sets (orange), the clusters (blue), and the control clusters (green). In addition, the most left bars represent the median over the 15 diseases.

Disease	Initial nb. of genes	Nb. of clusters	Clusters	Nb. of genes
Alzheimer	8	1	Cluster1	3
Anemia	44	1	Cluster1	4
Cardiomyopathy	22	1	Cluster1	6
Cataract	20	1	Cluster1	3
Colorectal cancer	21	1	Cluster1	4
Deafness	42	2	Cluster1	3
			Cluster2	3
Ehlers-danlos	10	1	Cluster1	5
Leukemia	112	9	Cluster1	8
			Cluster2	32
			Cluster3	12
			Cluster4	3
			Cluster5	5
			Cluster6	4
			Cluster7	3
			Cluster8	4
			Cluster9	5
Muscular dystrophy	24	2	Cluster1	3
			Cluster2	12
Myopathy	41	2	Cluster1	4
			Cluster2	4
Retinitis pigmentosa	30	2	Cluster1	7
			Cluster2	3
Spastic paraplegia	7	1	Cluster1	5
Spinocerebellar	7	1	Cluster1	4
Xeroderma pigmentosum	10	1	Cluster1	5
Zellweger	9	1	Cluster1	7

Tables

Table 1 - Clustering results

Results of clustering applied as a preprocessing method on our disease benchmark data. Only the 15 diseases with at least one valid cluster from the original set of 29 diseases are kept for further benchmark.

Chapter 6

Conclusion

The work presented in this thesis is focused on the gene prioritization problem. The concept of gene prioritization was introduced a decade ago and involves the use of computer power to sort a list of genes based on their characteristics with respect to a particular biological process. Most commonly, this biological process will be a disease and then, the highly ranked genes will be considered good candidates to pursue a biological validation.

Briefly, we review an exhaustive list of gene prioritization tools, we make a comparison of performance, we combine the strategies that have outperformed others in three different diseases and finally we develop a cluster analysis based method to enhance the accuracy of the gene rankings.

In our first study, a thorough comparison among free web-based gene prioritization tools is presented. A description of all of these is provided and a decision tree has been designed in order to help the final user to select the most appropriate tool for his data and his purposes. A list of publications where the different tools are validated and compared has been presented. This can give us a first, be it shallow, performance comparison between tools. The databases upon which the tools rely, have been categorized to allow the reader glimpsing which type of data needed by each of the tools when making comparisons and how eventually the ranking is constructed. All this information has been also uploaded to a website where we intend to keep up-to-date information about gene prioritization in a reference portal where the state of art of the field is presented.

In our second study, we have made a quantitative comparison among gene prioritization tools. It has been the first time that such a complete performance comparison has been performed and it has given us valuable information for our next project where the best performance tools have been selected to be combined in a holistic approach. To perform a comparison as fair as possible, we have run the tools on 42 genes published in certain biomedical journals as disease-genes within the 72 hours following their publication. This method allowed us to validate tools with an already known disease gene but the validation was done using tools that relied in databases still not updated with this information. We have applied this strategy to 42 newly discovered genes in a six months period and the results show differences of performance among the tools in terms both of accuracy of the final ranking and reliability.

Our third project takes advantage of the output of both the first and second study. We have combined the top performance tools of our previous work in a two-layer based strategy using order statistics and we have applied it to three different

diseases: Congenital heart disease (CHD), congenital diaphragmatic hernia (CDH) and asthma. The first layer includes the two best performance tools in our previous study when applied to the full genome. This first layer acts as a filter reducing the candidate set from the around 20.000 genes contained in the human genome contains to a set of 2000. This reduction of approximately 90% in the number of candidate genes is consistent with the results of our previous study. Order statistics is used to merge the different rankings of the first layer and a combined candidate set of 2000 genes is used in the second layer, where two other gene prioritization tools are used to reduce the list to a final ranking of 200 genes. Then, this final list is manually analyzed by experts to assess the validity of our approach. As we state in chapter 4, interesting results arose and novel interesting candidate genes are proposed for asthma.

Finally, in our fourth project, we propose a preprocessing step in gene prioritization using a transaction based cluster analysis strategy. CLOPE is a clustering algorithm that has been used in the last years for market basket type of data due to its histogram based utility function, which is a global measure that allows a fast evaluation of the goodness of a clustering distribution regardless of the amount of elements to be clustered. We used this study to show that clustering the training set generally leads to better rankings using different gene prioritization tools. We also showed that the gain in performance is not due to the reduction in size of the training sets by comparing the results of the real clusters with randomly built ones of equal size.

It is important to note that gene prioritization is not a goal in itself but a means to ease the discovery of new disease genes. Computer power can efficiently gather known information and incorporate unstructured data to rank possible candidate genes and present to the wet lab researcher an informative ranking of genes. However, a top ranked gene does not imply that the gene will surely be involved in the biological process of interest. Gene prioritization must be followed by a biological validation to ensure the involvement of the investigated gene in the disease.

6.1 OVERVIEW

In the first two studies, we have intended to shed light on the fast evolving gene prioritization field. We believe that the thorough analysis and description of gene prioritization tools of the first work and the subsequent performance study will help final users to select the most appropriate tool or set of tools for their needs.

We believe that our goal has been achieved because of two reasons. First, the seminal paper was published in a bioinformatics journal of high impact *Briefings in Bioinformatics* and after the publication of the review, less than two years ago, it has been cited in 35 different works. Second, the gene prioritization portal, designed to keep up-to-date information on this field has been extended with few other tools after communication with their developers, showing that the website can become a

reference in the gene prioritization field. Furthermore, since May 2010, when the site was launched, a total of 25565 visits have been registered (10737 unique visitors). Giving the specificity of the topic covered by the website, we believe that these figures are rather impressive.

With regard to the last two studies, their original goal was to work on the improvement of the already existing gene prioritization solutions. We have tried to achieve this by an integration of tools and a cluster analysis based preprocessing step. Validation experiments in chapter 5 show a general improvement in accuracy of the rankings, but estimating the quality of the results applied to real biological data is a more difficult task and deeply depends on the expertise of the users. Chapter 4 shows interesting results where top ranked genes (starting with the full genome) are related to similar conditions to the one described in the training set, or they have been confirmed in animal models, or they are involved in syndromic examples of the disease. These results are very encouraging since they show that using the largest candidate set possible (the full genome) is not an insurmountable obstacle to obtain a quality ranking.

The application of cluster analysis to gene prioritization has opened several questions, not all addressed in this work. Section 6.2 elaborates on these issues.

6.2 CLUSTER ANALYSIS AND GENE PRIORITIZATION

Choosing a training set is not an easy task and the complete process of gene prioritization depends on it. For non expert users, the selection of appropriate training genes is a hard and tedious task. Hard because limited expertise does often not allow those users to easily select or discard genes. A first approach for a non expert user would be to use publicly available databases on human genetics such as OMIM or GAD. However, this would not guarantee a complete training set due to incompleteness of these data sources. It is possible to complete the set using literature search engines like PubMed, but unfortunately this must be done manually turning the building of a training set into a tedious task.

The main objective that we pursue in our fourth study (chapter 5) is to ease the task of the final user by automatically selecting training sets. Even if the main goal has been accomplished and better rankings have been obtained, there remain open paths still to be explored. One of them that has attracted our attention consists of building a fully automatic high quality training set from scratch. We implemented an algorithm to select all the entries in OMIM related to a disease (neuropathy, in our experiment) and we applied the CLOPE clustering algorithm to this big and noisy training set in order to obtain high quality training sets. We found the AUC values obtained during validation to be inferior to those obtained by expert selected sets. We believe that the main reason for this failure was due to the limited amount of databases used. The relevance of this limitation was confirmed after several

meetings with biologists, geneticists and biochemists since some clusters obtained by CLOPE algorithm had no meaning in the eyes of the biologist. To overcome this situation, we believe that using additional databases and therefore extending the characteristics of the genes used during the clustering process would allow us to cope again with the fully automatic selection of a training set and would help us to find more not only *biologically meaningful* training sets.
Despite the fact that we have reviewed about 20 gene prioritization tools and that in the last years few more have been added to the gene prioritization portal, it is still uncommon to find tools that allow massive experiments. This is the reason why we only used two tools in our fourth study (Endeavour and Genedistiller). The huge amount of prioritizations run to validate 29 diseases and the 27 respective clusters can in no way be done manually and as long as other methods do not provide an API to automatically load thousands of runs, our conclusion about cluster analysis and gene prioritization will not be more general.
The clustering algorithm selected has been used during the last years in very similar problems to the one used in chapter 5. However, it leaves out of consideration several elements to be introduced in future studies. First, every gene will be only present in one cluster. However, we know that gene products, due to alternative splicing for instance, interact in different pathways and therefore can be related to different diseases. That is why we are developing a cluster analysis approach where genes can be distributed in different clusters and therefore can be present in different training sets.
Furthermore, CLOPE, like any other transactional clustering algorithm, takes only into account Boolean data. This has been a major limitation in our approach since many biological databases are not convertible to Boolean data. Our current efforts also take this into account to arrive at an algorithm for coping with different types of data.

6.3 OTHER LINES OF RESEARCH

6.3.1 HAEMATLAS

In addition to the results of the validation of cluster analysis to gene prioritization presented in chapter 5, we have been working with this approach on real life data.
For these project, we have repeated the strategy of clustering the training sets selected by the user as presented in chapter 5. The objective of this work is to propose new candidates related to platelet-based diseases, using hematopoiesis specific information. Hematopoiesis is the process leading to the differentiation of a common and undifferentiated cellular precursor called hematopoietic stem cell (HSC) into the diverse type of blood cells, which will eventually be distinct in number, shape and function.

To better understand hematopoiesis, in 2009 was launched the ambitious whole genome microarray based study called Haematlas [103], where gene expression profiles of human erythroblasts, megakaryocytes, B cells, cytotoxic and helper T cells, natural killer cells, granulocytes and monocytes were compared.

Out of this comprehensive analysis, we are particularly interested in megakaryocytes, since these are the cells responsible for the production of blood thrombocytes, more commonly known as platelets.

Thrombocytes, and by extension megakaryocytes, represent one of the different lines of research in *Latron*, the Laboratory for Thrombosis Research by Prof. Hans Deckmyn in KU Leuven campus KULAK.

Haematlas showed that 272 genes were lineage specific in megakaryocytes. With the aim of finding genes directly involved in platelet based diseases, we designed a gene prioritization experiment where the megakaryocyte lineage specific genes would become the candidate set to prioritize. Two training sets covering the most common types of platelet based diseases were built and a combined process of cluster analysis plus gene prioritization was launched. One of the training sets includes receptors with known function in platelets: Integrins (ITGA2, ITGB1, ITGA2B, ITGB3), G protein-coupled receptor family proteins (TXBA2R, AXL, MERTK, TYRO3, F2R, F2RL3, P2RY1, P2RY12, PTGIR, ADRA2A, PTAFR, AVPR1A, HTR2A), protein channels (ORAI, P2RX1, ITPR, RYR2) and other types of receptors (GPIBA, GPIBB, GPIX, GPV, GPVI). The second training set, included signal transduction genes (PLA2, PLCB2, PLCG2, PTGS1, TBXAS1, GNAQ, GNAS, Giα2, STIM, RASGRP1, FERMT3, APBB1IP and WAS.

Following the previous chapter, a preprocessing step based on cluster analysis with CLOPE, prior to gene prioritization with Endeavour, was applied for every experiment (one with receptors as training set and another one with signal transduction genes). A complete screen of all possible values of the repulsion parameter was launched and an optimum value of 4.9 was found for receptors and 2.1 for signal transduction genes.

The cluster analysis of the training set based on receptors, yielded two clusters of three and five genes respectively (F2RL3,P2RY1,P2RY12 and GPIBA,GPIBB,GPIX,GPV,GPVI). At first sight, both clusters seem reasonably homogeneous (five glycoproteins in one cluster and two purinergic receptors out of three genes). However, there are reasons to believe that other genes from the same families could belong to these clusters as well. We may not forget that the cluster analysis has been directed by a limited amount of data, limited both in number of databases –restricted by the type of data of the databases- and by incompleteness of databases (continuously growing in terms of quantity and quality of annotations).

The cluster analysis run on the training set of signal transduction genes returned a single cluster with three genes (RASGRP1, URP2, APBB1IP).

Three gene prioritization experiments with Endeavour have been launched using the 272 lineage-specific megakaryocytes genes as candidate set in the three of them

and the two clusters originated from receptor genes and the cluster from signal transduction genes as training set in each of them. We have run Endeavour with all databases available, and besides, we have added an external database to sharpen the accuracy of the prioritization. Haematlas data has been adapted to the requirements of Endeavour and the expression values of the genes both from the candidate set and from the training set have been taken into account during the prioritization process.

A thorough analysis of the rankings was performed by Dr. Katleen Broos from Latron department. Her analysis highlights two type of genes: those highly ranked and that have been linked to platelets in the last years (after the publication of Haematlas) and those which seem good candidates for further research. The first ones can be seen as a sort of validation of our rankings and encourage a further analysis on the second ones. Tables 6.1 and 6.2 show these two groups of genes when gene prioritization is performed on the two clusters from receptor genes and table 6.3 shows the analysis for the signal transduction cluster.

The three tables show a limited overlapping, more evident between receptor clusters. Giving that we have analyzed the top 50 genes in a total of 272, we expected some sort of overlapping. However, the position of the common highly ranked genes is not equal showing that the use of different training sets (even though similar) output different rankings. These results open new doors which should be explored. For instance, the ranking coming from clusters of the same training set could be fused in the spirit of the study of chapter 4.

Ranking	Gene	Description
Recently discovered		
4th	WNT11	Wnt signaling negatively regulates platelet function[104]
18th	ARMCX1	Regulated by Wnt pathway [105]
29th	ESAM	Inhibition of thrombus formation [106]
43rd	DCBLD2	Inhibition of thrombus formation [107]
Good Candidates		
10th	ATP2C1	Involved in Calcium transportation
17th	ARHGAP21	It codes for Rho GTPase activating protein and Rhoa mediator in platelet activation
50th	NLK	Wnt coupled

Table 6.1. Good candidates and genes recently discovered to be involved in platelets ranked in the top 50 when using receptors cluster 1 as a training set

Ranking	Gene	Description
Recently discovered		
5th	WNT11	Wnt signaling negatively regulates platelet function[104]
7th	LRRC32	Promotion of thrombus formation [107]
17th	ESAM	Regulates thrombus formation [106]
18th	DCBL2	Regulates thrombus formation [107]
36th	TSPAN9	Regulation of platelet function
Good Candidates		
21th	ARHGAP21	It codes for Rho GTPase activating protein and Rhoa mediator in platelet activation
32th	RSU11	Relation to platelets unknown but ras is involved in platelet activation
42th	PRICKLE2	Involved in wnt pathway
44th	SLC8A3	Involved in calcium transport

Table 6.2. Good candidates and genes recently discovered to be involved in platelets ranked in the top 50 when using receptors cluster 2 as a training set

Ranking	Gene	Description
Recently discovered		
26th	ESAM	Regulates thrombus formation [106]
30th	ARMCX1	Regulated by Wnt pathway [105]
44th	WNT11	Wnt signaling negatively regulates platelet function [104]
Good Candidates		
3rd	SPNS1	Linked to S1P with known function in platelets
4th	ARHGAP21	It codes for Rho GTPase activating protein and Rhoa mediator in platelet activation
13th	PLEK	Possible interactions in granule secretion, aggregation and actin polymerization
18th	NLK	Wnt coupled
41th	PRICKLE2	Involved in wnt pathway
43rd	ARHGAP18	It codes for Rho GTPase activating protein and Rhoa mediator in platelet activation

Table 6.3. Good candidates and genes recently discovered to be involved in platelets ranked in the top 50 when using signal transduction cluster 1 as a training set

6.3.2 *DAPHNIA* AND BICLUSTERING

Another project that has been undertaken during these last years of my research in Kortrijk is related to the application of bicluster analysis to transcriptomics data from *Daphnia magna*.

Daphnia, a water flea, has become in the last years a very interesting organism for several reasons. From an ecological perspective, this water flea is basic in aquatic ecosystems since it is both a primary grazer of algae and forage for fish. From an evolutionary and ecological point of view, it is a prototypical model organism as considerable evolutionary and ecological information about it is available (including relation against external agents, like host-parasite interaction). *Daphnia* also has a short generation time which makes it suitable for evolutionary studies and furthermore it allows to work with clonal lineages. In the last years, the amount of genomic data about *Daphnia* has been increasing, and an important milestone was reached with the sequencing of the full genome of *Daphnia pulex*, thanks to the joint work of the *Daphnia* Genomics Consortium (http://daphnia.cgb.indiana.edu).

A last reason of the imporante of *Daphnia*, is that as a crustacean, the water flea is expected to share a relatively high number of genes with other well known model organisms such as *Drosophila* or *Anopheles*.

We have been applying biclustering techniques to very recent transcriptomics data in order to find out groups of genes being expressed similarly in similar conditions. The reason of choosing bicluster analysis instead of cluster analysis is basically due to the fact that we aim at seeking genes that do not necessarily share a common pattern of expression through all the conditions of the experiment. Based on a benchmark [108], we have used OPSM [109], Samba [110] and ISA [111]. The work is still in progress and several biclusters of a size manually manageable have been found and are currently being analyzed.

APPENDIX A
GENE RANKINGS

1	TGFBR2	41	MYH10	81	MMP2	121	INSR	161	NPPA
2	BMPR1A	42	CASQ2	82	JUN	122	E2F1	162	RUNX2
3	NOTCH2	43	MITF	83	TNNI3	123	ESR2	163	GAS6
4	NOTCH3	44	ACTA1	84	MYL7	124	ACTN2	164	PDX1
5	TGFBR1	45	ITGB1	85	TP63	125	SLC25A4	165	GSN
6	PDGFRB	46	FBLN1	86	EDNRA	126	PSEN1	166	MYLK
7	FBN1	47	GJA5	87	PITX2	127	FLT1	167	IRS1
8	MYH7	48	BRAF	88	CAV1	128	RELA	168	SRF
9	LMNA	49	TGFB2	89	ERBB2	129	HAND2	169	YY1
10	FGFR1	50	HAND1	90	LMX1B	130	PTK2	170	EPAS1
11	TGFB1	51	NKX2-1	91	ITGAV	131	BMPR2	171	ERBB4
12	MYBPC3	52	TEK	92	RET	132	COL3A1	172	NID1
13	ACVR1	53	BMP4	93	FN1	133	INHBC	173	EGR1
14	CREBBP	54	DES	94	MYH9	134	TWIST1	174	VEGFC
15	KIT	55	TBX3	95	FHL2	135	TGFB3	175	TFAP2C
16	LAMC1	56	RB1	96	MECOM	136	FLNA	176	LEF1
17	TNNT2	57	EFEMP1	97	TNFRSF1A	137	STAT1	177	NOG
18	FGFR2	58	MET	98	SOX9	138	GATA2	178	FLT3
19	PPARG	59	FGFR3	99	MSX1	139	IL9R	179	ITGB3
20	GLI3	60	IGF1R	100	SMAD2	140	VDR	180	SOX10
21	ACVRL1	61	TTN	101	SMAD4	141	GDF5	181	NF1
22	HNF1A	62	AKT1	102	CAV3	142	GATA3	182	HNF4A
23	CTNNB1	63	BMP2	103	ACTB	143	TGIF1	183	PDGFA
24	EP300	64	KDR	104	HIF1A	144	EPHB2	184	PPARD
25	COL1A1	65	KRAS	105	SMAD3	145	HRAS	185	TDGF1
26	PAX3	66	INHBA	106	TBX2	146	MAP2K1	186	GDNF
27	DLL1	67	FOXO1	107	LAMA5	147	COL1A2	187	MAPKAPK3
28	EGFR	68	SP1	108	IKBKB	148	ITGB4	188	RPS6KA3
29	GTF2I	69	MYL2	109	CDX2	149	AR	189	MUSK
30	ACVR1B	70	TP53	110	IGF1	150	SOX2	190	CRYAB
31	ACVR2A	71	CSRP3	111	NRP2	151	MYL3	191	MAPK14
32	TFAP2A	72	WT1	112	THBS1	152	RARG	192	CDK4
33	HSPG2	73	MYC	113	ACAN	153	NOS3	193	BMP10
34	FLT4	74	BMP7	114	ENG	154	THRB	194	SP3
35	GLI2	75	ILK	115	PLAT	155	PAX6	195	GJA3
36	GTF2IRD1	76	LTBP1	116	APC	156	PDGFB	196	CRIP2
37	FBN2	77	PTPN11	117	ABL1	157	AXL	197	BMPR1B
38	SHH	78	NRP1	118	STAT3	158	AKT2	198	SMAD6
39	TNNC1	79	VEGFA	119	NFKB1	159	ERBB3	199	THBS4
40	PDGFC	80	PBX1	120	GDF11	160	PTEN	200	CUX1

Table A.1. – Top 200 genes in congenital heart disease

Final rankings for congenital heart disease obtained combining Pinta, Candid, Endeavour and GeneDistiller in the two-layer gene prioritization strategy.

1	NR1H3	41	MET	81	STAT3	121	NR4A1	161	ALDH3A1
2	PPARG	42	RELA	82	NOTCH1	122	GRHPR	162	ZBTB16
3	PPARA	43	NR2E3	83	ALDH7A1	123	NOTCH2	163	MITF
4	ALDH2	44	GATA6	84	FGFR1	124	SMAD2	164	GJA1
5	AR	45	CTNNB1	85	CCND1	125	EHHADH	165	RPE65
6	THRB	46	HSD17B8	86	HAND1	126	PLAT	166	NR2C1
7	ESR2	47	CDK4	87	ALDH1L1	127	ECHS1	167	NF1
8	NR1I3	48	MYC	88	EFEMP1	128	ALDH3B2	168	CASP3
9	CYP1A2	49	ALDH3A2	89	PGR	129	WWOX	169	ALDH6A1
10	THRA	50	TBX5	90	CYP3A4	130	UGT1A5|UGT1A6	170	HIF1A
11	HNF4A	51	GLI2	91	SP1	131	CYP7A1	171	NID1
12	NR1I2	52	FABP1	92	CYP3A5	132	PDGFRB	172	ARNT
13	NKX2-5	53	NR3C1	93	ITGB1	133	HAO1	173	ALB
14	ESR1	54	ALDH9A1	94	APAF1	134	PSEN1	174	STK11
15	HSD11B1	55	DHRS4	95	E2F1	135	NR0B1	175	HSD3B7
16	GRK1	56	CYP4A22	96	SOD1	136	ALDH4A1	176	ABCG5
17	CYP2E1	57	NFKB1	97	NR5A2	137	EPHX2	177	HSD11B2
18	CREBBP	58	AKT1	98	RORC|LINGO4	138	BMP4	178	MECOM
19	NR0B2	59	CYP2B6	99	FASN	139	KIT	179	MMP2
20	NR1H2	60	SHH	100	CYP1A1	140	SOX2	180	FGG
21	PPARD	61	APOA1	101	COL1A1	141	CYP19A1	181	FGFR4
22	NR1H4	62	EP300	102	GJB1	142	TH	182	ALDH3B1
23	NR2F6	63	PLG	103	NR5A1	143	FGFR2	183	KLF5
24	RLBP1	64	SOX9	104	BMPR1A	144	PAX3	184	VTN
25	HSD17B6	65	RB1	105	PROX1	145	MAOB	185	TGFBR1
26	ALDH8A1	66	HSD17B2	106	TTR	146	VEGFA	186	IGF2
27	VDR	67	ACAA1	107	LMNA	147	RET	187	CEBPB
28	HNF1A	68	EGFR	108	TNFRSF1A	148	NOS2	188	GATA3
29	BDH1	69	DHRS1	109	AKR1C4	149	NOS1	189	BMP2
30	BRAF	70	TP63	110	COMT	150	PECR	190	TCF3
31	NR2F1	71	JAG1	111	ALDH1B1	151	KRAS	191	LPL
32	GLI3	72	TGFBR2	112	CEBPA	152	ACOX1	192	PTHLH
33	POR	73	PTGS2	113	SMAD3	153	PBX1	193	GAPDH
34	TP53	74	CYP4F2	114	PDGFRA	154	TGIF1	194	IRS1
35	ESRRA	75	IGF1	115	NR6A1	155	NFE2L2	195	NR2C2
36	LRP2	76	NR4A2	116	WT1	156	CAV1	196	CYP11A1
37	XDH	77	HSD17B4	117	GATA2	157	ITGAV	197	RORB
38	CYP4A11	78	APOE	118	IGF1R	158	PTPN11	198	NSD1
39	ASMT	79	ERBB2	119	TGFB1	159	HSD17B12	199	NME1|NME2
40	CYP27A1	80	AKT2	120	CDKN1A	160	NCOR2	200	TGM2

Table A.2. – Top 200 genes in congenital diaphragmatic hernia

Final rankings for congenital diaphragmatic hernia obtained combining Pinta, Candid, Endeavour and GeneDistiller in the two-layer gene prioritization strategy.

1	TNF	41	MAPK14	81	KRAS	121	FGFR3	161	SMAD1
2	TGFB1	42	MYC	82	PTPN11	122	CXCR3	162	PLG
3	IL1B	43	IFNGR2	83	PTGS2	123	JUN	163	KDR
4	PTPRC	44	IL6ST	84	MMP3	124	ICAM1	164	HRAS
5	TGFBR2	45	FASLG	85	TLR7	125	MET	165	NOTCH2
6	TLR4	46	TLR5	86	IL27RA	126	CCL3	166	FBN1
7	MYD88	47	PDGFRA	87	PTPN6	127	FLT3	167	ITGAV
8	IL6	48	THBS1	88	PPARG	128	BCL2	168	MMP1
9	SMAD2	49	PLAT	89	NR3C1	129	CSF3R	169	PIK3R1
10	IL1RAP	50	MAP2K1	90	CEBPB	130	RXRA	170	IL15
11	TLR2	51	RARA	91	CASP8	131	CD80	171	HNF4A
12	NFKB1	52	IL3RA	92	CD40	132	TNFSF10	172	HGF
13	CSF2RA	53	SMAD4	93	IL1A	133	VCAN	173	IL2RB
14	INPP5D	54	CCL5	94	STAT1	134	ITGA4	174	TLR8
15	CD4	55	IL7R	95	SERPINE1	135	ETS1	175	NFKB2
16	IL10RA	56	ADAM9	96	IL6R	136	SELL	176	CCL24
17	RELA	57	PDGFRB	97	CD44	137	LEPR	177	PECAM1
18	CCL2	58	FGFR1	98	IL1RL2	138	HIF1A	178	ACAN
19	FAS	59	ITGB1	99	IKBKB	139	HLA-B	179	IFNAR2
20	IKZF1	60	CCR5	100	FGFR2	140	ESR2	180	IL1RN
21	CREBBP	61	MMP14	101	TRAF6	141	ERBB2	181	FLNA
22	IL2RG	62	CD40LG	102	ALOX5	142	IRF1	182	ITGB3
23	ITGB2	63	ADAM17	103	CCR2	143	C6	183	SMAD7
24	IL12B	64	PSEN1	104	FLT4	144	CX3CL1	184	MMP13
25	C3	65	CCR1	105	CASP3	145	FOS	185	TGFB3
26	TNFRSF1A	66	EGFR	106	CSF1	146	NCOA3	186	ACTB
27	IL18RAP	67	FCER1G	107	PIK3CA	147	CCR7	187	APC
28	IL2	68	IKBKG	108	GHR	148	PTPN22	188	THBD
29	CD86	69	TGFB2	109	CTNNB1	149	JAK2	189	MERTK
30	FN1	70	FCGR2A	110	CSF2	150	VAV1	190	ATM
31	TP53	71	IL9R	111	ITGAX	151	EFEMP1	191	WAS
32	AKT1	72	STAT6	112	IL5	152	AXIN1	192	PDGFB
33	NFKBIA	73	RB1	113	KIT	153	TLR3	193	HLA-C
34	CSF2RB	74	CSF1R	114	TLR6	154	CCL7	194	AC046176.1|LYN
35	IL21R	75	IL10RB	115	RAF1	155	ITGAL	195	IFNAR1
36	TNFRSF1B	76	MMP9	116	PIK3CG	156	ADAM10	196	HBEGF
37	FCGR2B	77	BMP2	117	CXCR2	157	CASP1	197	KITLG
38	IFNGR1	78	TLR1	118	PTPRF	158	IL13RA1	198	FCGR1A
39	CXCR4	79	MAPK1	119	COL1A1	159	HLA-A|HLA-G	199	IL12A
40	STAT3	80	EP300	120	C3AR1	160	PTPRJ	200	PRKCD

Table A.3. – Top 200 genes in asthma

Final rankings for asthma obtained combining Pinta, Candid, Endeavour and GeneDistiller in the two-layer gene prioritization strategy.

APPENDIX B
CLUSTERING AND GENE PRIORITIZATION

Disease	***Repulsion***
Alzheimer's disease	4.4
Amyotrophic lateral sclerosis	2.0
Anemia	12.8
Breast cancer	11.0
Cardiomyopathy	5.5
Cataract	13.0
Charcot-Marie-Tooth disease	7.5
Colorectal cancer	7.2
Deafness	19.4
Diabetes	10.7
Dystonia	2.0
Ehlers-Danlos syndrome	2.7
Epilepsy	5.1
Hemolytic anemia	4.4
Ichthyosis	4.7
Leukemia	11.8
Lymphoma	20.0
Mental retardation	10.7
Muscular dystrophy	5.2
Myopathy	9.8
Neuropathy	6.8
Obesity	5.7
Parkinson's disease	3.6
Retinitis pigmentosa	6.8
Spastic Paraplegia	2.3
Spinocerebellar ataxia	3.9
Usher syndrome	10.2
Xeroderma Pigmentosum	3.9
Zellweger Syndrome	2.7

Table B.1: Optimum repulsion (r) parameter for every disease in CLOPE. R controls the tightness of the clusters and its value changes according to the sparseness of the space solution.

Disease	***Genes***	
alzheimer	Cluster 1	BPTF BLMH ACE
anemia	Cluster 1	FANCD2 FANCE FANCF FANCC
cardiomyopathy	Cluster 1	MYL2 TCAP TNNC1 LMNA DES ACTC1
cataract	Cluster 1	CRYGC CRYGD CRYBA1
colorectal_cancer	Cluster 1	MLH1 PMS2 MLH3 PMS1
deafness	Cluster 1	KIAA1199 COCH ENSG00000166763
	Cluster 2	TMC1 ACTG1 GJB2
ehlers-danlos	Cluster 1	COL5A1 COL1A1 COL3A1 COL1A2 COL2A1
emolytic_anemia	Cluster 1	BPGM G6PD GPI TPI1
leukemia	Cluster 1	ARHGEF12 TCL1A RALA ARHGAP26 SH3GL1 BCR MLLT4 CHIC2
	Cluster 2	GATA1 MLF2 MLLT6 PSMD7 TAL1 HLF TAL2 MLLT1 LMO1

		TCTA FOXN2 MLLT3 HOXA9 TET1 C1ORF56 LMO2 LPP ETV6 LYL1 TCL6 DLEU2L BAALC ELL MLLT10 MLL5 DLEU1 FLI1 RBM15 AFF1 MLF1 RAP1GDS1 ZMYM2
	Cluster 3	PBX3 IKZF1 ERG MLL2 WHSC1L1 CBFB MLL4 TCF3 RARA DEK MKL1 MLL3
	Cluster 4	P2RX7 LIFR MPL
	Cluster 5	FUS SLC20A2 SLC20A1 MME KDSR
	Cluster 6	ERBB4 KIT FLT3 PDGFRB
	Cluster 7	PBX1 IRF1 ARNT
	Cluster 8	TLX3 THRB TLX1 TLX2
	Cluster 9	PICALM

		SEPT9 NUMA1 ACSL1 FNBP1
muscular_dystrophy	Cluster 1	FRG1 LMNA PABPN1
	Cluster 2	SEPN1 PLEC DYSF FKTN SGCG EMD MYOT SGCD SGCB FKRP TCAP SGCA
Myopathy	Cluster 1	TNNC1 MYL2 MYH2 ACTC1
	Cluster 2	COL6A3 COL9A3 COL6A2 COL6A1
retinitis_pigmentosa	Cluster 1	ROM1 PRPH2 RP1 TULP1 PDE6B PDE6A RPGRIP1
	Cluster 2	RP9 PRPF31 PRPF8
spastic_paraplegia	Cluster 1	IPA1 SPAST GPM6B ATL1 BSCL2
spinocerebellar_ataxia	Cluster 1	ATXN7 TBP TDP1 ATXN1
xeroderma_pigmentosum	Cluster 1	DDB2 XPC DDB1 ERCC5 ERCC4
zellweger_syndrome	Cluster 1	PEX5 PEX16 PEX13 PEX10

		PEX2 PEX3 PEX1

Table B.2: Clusters with three or more elements obtained from gene sets on Supplementary Table 2 using CLOPE with r value obtained from Supplementary Table 1.

Syndrome	Full gene set	Genes in clusters
Alzheimer's disease	0,11745	0,113066667
Anemia	0,0653	0,0228
Cardiomyopathy	0,038818182	0,0166
Cataract	0,04966	0,010133333
Colorectal cancer	0,091238095	0,0101
Deafness	0,148142857	0,1492
Ehlers Danlos Syndrome	0,0364	0,0102
Leukemia	0,100121429	0,0952
Muscular distrophy	0,025316667	0,04336
Myopathy	0,062770732	0,01075
Retinitis Pigmentosa	0,056706667	0,01624
Spastic Paraplegia	0,047885714	0,01208
Spinocerebellar ataxia	0,047828571	0,0698
Xeroderma Pigmentosum	0,022	0,0216
Zellweger Syndrome	0,026444444	0,013371429

Table B.3: Comparison of AAC between training sets and clustered genes in Endeavour. Diseases which return more than one cluster show the average AAC value of all of them.

Syndrome	Full gene set	Genes in clusters
Alzheimer's disease	0,0437	0,45586664
Anemia	0,04657136	0,0086
Cardiomyopathy	0,05733332	0,01840004
Cataract	0,01355792	0,01146672
Colorectal cancer	0,10881904	0,0113
Deafness	0,06289	0,01716662
Ehlers Danlos Syndrome	0,03852	0,02744
Leukemia	0,08589052	0,110887879
Muscular distrophy	0,03156368	0,041418656
Myopathy	0,06196756	0,04635
Retinitis Pigmentosa	0,02681372	0,010080008
Spastic Paraplegia	0,09746672	0,30373328
Spinocerebellar ataxia	0,03594288	0,0375
Xeroderma Pigmentosum	0,01668	0,01248
Zellweger Syndrome	0,04865	0,07133332

Table B.4: Comparison of AAC between training sets and clustered genes in Genedistiller. Diseases which return more than one cluster show the average AAC value of all of them.

Syndrome	Full TS	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	Cluster8	Cluster9
Deafness	0,1492	0,078933333	0,219467							
Leukemia	0,0952	0,1373	0,106563	0,035233	0,040133	0,32296	0,0115	0,078533	0,0263	0,0364
Muscular distrophy	0,0434	0,067466667	0,037333							
Myopathy	0,0107	0,0114	0,0101							
Retinitis Pigmentosa	0,0162	0,0108	0,028933							

Table B.5: Comparison of AAC between training sets and individual clusters (for diseases with more than one cluster) using Endeavour

Syndrome	Full TS	Cluster1	Cluster2	Cluster3	Cluster4	Cluster 5	Cluster 6	Cluster7	Cluster 8	Cluster9
Deafness	0,05025	0,0062	0,028133							
Leukemia	0,06546792	0,14275	0,138193	0,0717	0,0636	0,0813	0,0319	0,038133	0,0598	0,18488
Muscular distrophy	0,01687264	0,165333	0,01044							
Myopathy	0,03277832	0,0323	0,0604							
Retinitis Pigmentosa	0,01382072	0,010343	0,009467							

Table B.6: Comparison of AAC between training sets and individual clusters (for diseases with more than one cluster) using Genedistiller

	ENDEAVOUR			GENEDISTILLER		
Disease	Random AAC	Cluster AAC	Difference	Random AAC	Cluster AAC	Difference
Alzheimer's disease	0,1274	0,1131	0,0144	0,00805332	0,06693332	-0,0589
Anemia	0,1347	0,0228	0,1119	0,04272376	0	0,0427
Cardiomyopathy	0,0462	0,0166	0,0296	0,00676604	0,00833324	-0,0016
Cataract	0,1608	0,0101	0,1507	0,01094664	0,00720008	0,0037
Colorectal cancer	0,1149	0,0101	0,1048	0,01731	0,0028	0,0145
Deafness Cluster 1	0,1586	0,0789	0,0797	0,07326664	0,0008	0,0725
Deafness cluster 2	0,1920	0,2195	-0,0275	0,03182072	0,00439996	0,0274
Ehlers Danlos	0,0256	0,0102	0,0155	0,010184	0,0136	-0,0034

Syndrome						
Leukemia Cluster 1	0,1291	0,1373	-0,0082	0,0922	0,0527	0,0395
Leukemia Cluster 2	0,1190	0,1066	0,0124	0,07212964	0,12528268	-0,0532
Leukemia Cluster 3	0,1179	0,0352	0,0827	0,07657088	0,02736668	0,0492
Leukemia Cluster 4	0,2478	0,0401	0,2077	0,13066148	0,01746672	0,1132
Leukemia Cluster 5	0,1013	0,3230	-0,2216	0,068176	0,0488	0,0194
Leukemia Cluster 6	0,2140	0,0115	0,2025	0,12347	0,0101	0,1134
Leukemia Cluster 7	0,2025	0,0785	0,1240	0,09148764	0,01946672	0,0720
Leukemia Cluster 8	0,1593	0,0263	0,1330	0,10254444	0,0205	0,0820
Leukemia Cluster 9	0,1652	0,0364	0,1288	0,09653096	0,33488	-0,2383
Muscular distrophy Cluster 1	0,0733	0,0675	0,0059	0,04377924	0,00933328	0,0344
Muscular distrophy Cluster 2	0,0391	0,0373	0,0017	0,05543468	0,00232	0,0531
Myopathy Cluster 1	0,1095	0,0114	0,0981	0,01904436	0,0074	0,0116
Myopathy Cluster 2	0,0776	0,0101	0,0675	0,0288	0,0034	0,0254
Retinitis Pigmentosa Cluster 1	0,0804	0,0108	0,0696	0,00593524	0,00217144	0,0038
Retinitis Pigmentosa Cluster 2	0,0449	0,0289	0,0160	0,00900004	0,00026668	0,0087
Spastic Paraplegia	0,0445	0,0121	0,0324	0,0253954	0,03933332	-0,0139
Spinocerebellar ataxia	0,0528	0,0698	-0,0170	0,01787	0,038	-0,0201
Xeroderma Pigmentosum	0,0408	0,0216	0,0192	0,004832	0,00232	0,0025
Zellweger Syndrome	0,0313	0,0134	0,0179	0,0016524	0,00046672	0,0012

Table B.7: Comparison of AACs between valid and random clusters using Endeavour and Genedistiller. The random cluster AAC values are normally distributed. This property is used to derive a p-value for the real cluster AAC using the normal Cumulative Distribution Function (CDF).

Bibliography

1. Brookes AJ: **The essence of SNPs**. *Gene* 1999, **234**:177–186.

2. Stankiewicz P, Lupski JR: **Structural variation in the human genome and its role in disease**. *Annu. Rev. Med.* 2010, **61**:437–455.

3. Faderl S, Talpaz M, Estrov Z, Kantarjian HM: **Chronic myelogenous leukemia: biology and therapy**. *Ann. Intern. Med.* 1999, **131**:207–219.

4. Vardiman JW, Harris NL, Brunning RD: **The World Health Organization (WHO) classification of the myeloid neoplasms**. *Blood* 2002, **100**:2292–2302.

5. Adams F: *Genuine Works of Hippocrates*. Krieger Pub Co; 1972.

6. Correns C: *Untersuchungen über die Xenien bei Zea Mays*. Borntraeger; 1899.

7. De Vries H: **Sur la loie de disjonction des hybrides**. *Comptes Rendus de l'Academie des Sciences Paris* 1900, **130**:845–847.

8. Bearn AG: **Inborn errors of metabolism: Garrod's legacy.** *Mol Med* 1996, **2**:271–273.

9. **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**:931–945.

10. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. *Nucleic Acids Res* 2005, **33**:D34–D38.

11. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structures**. *J. Mol. Biol.* 1977, **112**:535–542.

12. Dayhoff M, Eck R, Chang M, Sochard M: *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring; 1965, **1**.

13. McKusick VA: *Mendelian inheritance in man: catalogs of autosomal dominant, autosomal recessive and X-linked phenotypes*. Johns Hopkins University Press, Baltimore; 1966.

14. Galperin MY, Fernandez-Suarez XM: **The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection**. *Nucleic Acids Research* 2011, **40**:D1–D8.

15. Jorde LB: **Linkage disequilibrium as a gene-mapping tool.** *Am J Hum Genet* 1995, **56**:11–14.

16. Hardy J, Singleton A: **Genomewide association studies and human disease**. *N. Engl. J. Med* 2009, **360**:1759–1768.

17. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining**. *Nat. Genet* 2002, **31**:316–319.

18. Zhang P, Zhang J, Sheng H, Russo JJ, Osborne B, Buetow K: **Gene functional similarity search tool (GFSST)**. *BMC Bioinformatics* 2006, **7**:135.

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

20. Hutz JE, Kraja AT, McLeod, HL, Province MA: **CANDID: a flexible method for prioritizing candidate genes for complex human traits**. *Genet. Epidemiol.* 2008, **32**:779–790.

21. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y: **Gene prioritization through genomic data fusion**. *Nat Biotech* 2006, **24**:537–544.

22. Chen J, Xu H, Aronow B, Jegga A: **Improved human disease candidate gene prioritization using mouse phenotype**. *BMC Bioinformatics* 2007, **8**:392.

23. Yue P, Melamud E, Moult J: **SNPs3D: Candidate gene and SNP selection for association studies**. *BMC Bioinformatics* 2006, **7**:166.

24. Seelow D, Schwarz JM, Schuelke M: **GeneDistiller—Distilling Candidate Genes from Linkage Intervals**. *PLoS ONE* 2008, **3**:e3874.

25. Yoshida Y, Makita Y, Heida N, Asano S, Matsushima A, Ishii M, Mochizuki Y, Masuya H, Wakana S, Kobayashi N, Toyoda T: **PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning**. *Nucleic Acids Res* 2009, **37**:W147–152.

26. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S: **TOM: a web-based integrated approach for identification of candidate disease genes**. *Nucleic Acids Research* 2006, **34**:W285–W292.

27. Masotti D, Nardini C, Rossi S, Bonora E, Romeo G, Volinia S, Benini L: **TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders**. *Bioinformatics* 2008, **24**:428–429.

28. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS: **PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites**. *Nucleic Acids Research* 2008, **36**:W399–W405.

29. Driel MA van, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM: **A text-mining analysis of the human phenome**. *European Journal of Human Genetics* 2006, **14**:535–542.

30. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM: **Using literature-based discovery to identify disease candidate genes**. *Int J Med Inform* 2005, **74**:289–298.

31. Van Vooren S, Thienpont B, Menten B, Speleman F, De Moor B, Vermeesch J, Moreau Y: **Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations**. *Nucleic Acids Res* 2007, **35**:2533–2543.

32. Yu W, Wulf A, Liu T, Khoury M, Gwinn M: **Gene Prospector: An evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases**. *BMC Bioinformatics* 2008, **9**:528.

33. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA: **Analysis of protein sequence and interaction data for candidate disease gene prediction**. *Nucleic Acids Res* 2006, **34**:e130.

34. Xiong Q, Qiu Y, Gu W: **PGMapper: a web-based tool linking phenotype to genes**. *Bioinformatics* 2008, **24**:1011 –1013.

35. Kohler S, Bauer S, Horn D, Robinson P: **Walking the Interactome for Prioritization of Candidate Disease Genes**. *The American Journal of Human Genetics* 2008, **82**:949–958.

36. Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, Mooney SD: **An integrated approach to inferring gene–disease associations in humans**. *Proteins: Structure, Function, and Bioinformatics* 2008, **72**:1030–1037.

37. Gaulton KJ, Mohlke KL, Vision TJ: **A computational system to select candidate genes for complex human traits**. *Bioinformatics* 2007, **23**:1132–1140.

38. Morrison JL, Breitling R, Higham DJ, Gilbert DR: **GeneRank: Using search engine technology for the analysis of microarray experiments**. *BMC Bioinformatics* 2005, **6**:233.

39. Ma X, Lee H, Wang L, Sun F: **CGI: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data**. *Bioinformatics* 2007, **23**:215–221.

40. Braun TA, Shankar SP, Davis S, O'Leary B, Scheetz TE, Clark AF, Sheffield VC, Casavant TL, Stone EM: **Prioritizing regions of candidate genes for efficient mutation screening**. *Hum. Mutat.* 2006, **27**:195–200.

41. Turner FS, Clutterbuck DR, Semple CAM: **POCUS: mining genomic sequence annotation to predict disease genes**. *Genome Biol.* 2003, **4**:R75.

42. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes**. *Am. J. Hum. Genet.* 2006, **78**:1011–1025.

43. Tiffin N, Okpechi I, Perez-Iratxeta C, Andrade-Navarro MA, Ramesar R: **Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes**. *Physiol. Genomics* 2008, **35**:55–64.

44. Nitsch D, Tranchevent L-C, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y: **Network analysis of differential expression for the identification of disease-causing genes**. *PLoS ONE* 2009, **4**:e5526.

45. Tranchevent L-C, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y: **ENDEAVOUR update: a web resource for gene prioritization in multiple species**. *Nucleic Acids Research* 2008, **36**:W377–W384.

46. Perez-Iratxeta C, Bork P, Andrade-Navarro MA: **Update of the G2D tool for prioritization of gene candidates to inherited diseases**. *Nucleic Acids Research* 2007, **35**:W212–W216.

47. Smith NGC, Eyre-Walker A: **Human disease genes: patterns and predictions**. *Gene* 2003, **318**:169–175.

48. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L: **The human disease network**. *Proc. Natl. Acad. Sci. U.S.A.* 2007, **104**:8685–8690.

49. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes**. *Nature* 2001, **409**:853–855.

50. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL: **Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA**. *Science* 1989, **245**:1066–1073.

51. Macdonald M, Ambrose C, Duyao M, Myers R, Lin C, Lakshmi S, Barnes G, Taylor S, James M, Groot N, MacFarlane H, Jenkins B, Anderson MA, Wexler N, Gusella J: **A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group**. *Cell* 1993, **72**:971–983.

52. Hunter DJ, Kraft P: **Drinking from the fire hose--statistical issues in genomewide association studies**. *N. Engl. J. Med.* 2007, **357**:436–439.

53. Jain AK, Mao RPW.: **Statistical pattern recognition: A review**. *IEEE Transactions on pattern analysis and machine intelligence* 2000, **22**:4–37.

54. Berkhin P: **A survey of clustering data mining techniques**. *Grouping Multidimensional Data* 2006:25–71.

55. Fisher DH: **Knowledge acquisition via incremental conceptual clustering**. *Machine learning* 1987, **2**:139–172.

56. Charikar M, Chekuri C, Feder T, Motwani R: **Incremental clustering and dynamic information retrieval**. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*. 1997:626–635.

57. Baldi P, Hatfield GW: *DNA microarrays and gene expression: from experiments to data analysis and modeling*. Cambridge University Press; 2002.

58. Shi J, Malik J: **Normalized cuts and image segmentation**. *IEEE Transactions on pattern analysis and machine intelligence* 2000, **22**:888–905.

59. Arabie P, Hubert L: **Cluster Analysis in Marketing Research**. *Advanced Methods in Marketing Research* 1994:160–189.

60. Portnoy L, Eskin E, Stolfo S: **Intrusion detection with unlabeled data using clustering**. In *Proceedings of ACM CSS Workshop on Data Mining Applied to Security*. 2001.

61. Phua C, Lee V, Smith K, Gayler R: **A comprehensive survey of Data Mining-based Fraud Detection Research**. *Artificial Intelligence Review* 2005.

62. Jain AK: **Data clustering: 50 years beyond K-means**. *Pattern Recognition Letters* 2010, **31**:651–666.

63. Anderberg MR: *Cluster Analysis for Applications*. New York: Academic Press; 1973.

64. Han J, Kamber M: *Data Mining: Concepts and Techniques*. 1st edition. Morgan Kaufmann; 2000.

65. Steinhaus H: **Sur la division des corp materiels en parties**. *Bull. Acad. Polon. Sci* 1956, **1**:801–804.

66. Hartigan JA, Wong MA: **Algorithm AS 136: A K-Means Clustering Algorithm**. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1979, **28**:100–108.

67. Kaufman L, Rousseeuw PJ: *Finding Groups in Data: An Introduction to Cluster Analysis*. 9th edition. Wiley-Interscience; 1990.

68. Ng RT, Han J: **Efficient and effective clustering methods for spatial data mining**. In *Proceedings of the International Conference on Very Large Data Bases*. 1994:144–144.

69. Guha S, Rastogi R, Shim K: **CURE: An Efficient Clustering Algorithm for Large Databases**. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*. 1998:73–84.

70. Zhang T, Ramakrishnan R, Livny M: **BIRCH: an efficient data clustering method for very large databases**. *ACM SIGMOD Record* 1996, **25**:103–114.

71. Ester M, Kriegel HP, Sander J, Xu X: **A density-based algorithm for discovering clusters in large spatial databases with noise**. In *Proc. KDD*. 1996, **96**:226–231.

72. Ankerst M, Breunig MM, Kriegel HP, Sander J: **OPTICS: Ordering points to identify the clustering structure**. *ACM SIGMOD Record* 1999, **28**:60.

73. Hinneburg A, Keim DA: **An efficient approach to clustering in large multimedia databases with noise**. *Knowledge Discovery and Data Mining* 1998, **5865**.

74. Wang W, Yang J, Muntz R: **STING: A statistical information grid approach to spatial data mining**. In *Proceedings of the International Conference on Very Large Data Bases*. 1997:186–195.

75. Sheikholeslami G, Chatterjee S, Zhang A: **Wavecluster: A multi-resolution clustering approach for very large spatial databases**. In *Proceedings of the International Conference on Very Large Data Bases*. 1998:428–439.

76. Agrawal R, Gehrke J, Gunopoulos D, Raghavan P: **Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications**. *ACMSIMOD* 1998, **72**:94–105.

77. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via de EM Algorithm**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1977, **39**:1–38.

78. Michalski RS, Stepp R: **Automated Construction Of Classifications Conceptual Clustering Versus Numerical Taxonomy**. 1983.

79. Huang Z: **Extensions to the k-means algorithm for clustering large data sets with categorical values**. *Data Mining and Knowledge Discovery* 1998, **2**:283–304.

80. Aranganayagi S, Thangavel K: **Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure**. *International Journal of Computational Intelligence* 2009, **5**:182–190.

81. He Z, Xu X, Deng S, Dong B: **K-Histograms: An Efficient Clustering Algorithm for Categorical Dataset**. *cs/0509033* 2005.

82. Guha S, Rastogi R, Shim K: **Rock: A robust clustering algorithm for categorical attributes* 1**. *Information Systems* 2000, **25**:345–366.

83. Gupta GK, Ghosh J: **Value Balanced Agglomerative Connectivity Clustering**. *IN SPIE PROC* 2001, **4384**:6–15.

84. Dutta M, Mahanta AK, Pujari AK: **QROCK: A Quick Version of the ROCK Algorithm for Clustering of Categorical Data**. 2005.

85. Jin Y, Zuo W: **Clustering Categorical Data Using Qualified Nearest Neighbors Selection Model**. *AI 2006: Advances in Artificial Intelligence* 2006:1037–1041.

86. Peters M, Zaki MJ: **CLICK: Clustering categorical data using k-partite maximal cliques**. In *IEEE International Conference on Data Engineering*. 2005.

87. Gibson D, Kleinberg J, Raghavan P: **Clustering categorical data: An approach based on dynamical systems**. *The VLDB Journal—The International Journal on Very Large Data Bases* 2000, **8**:236.

88. Ganti V, Gehrke J, Ramakrishnan R: **CACTUS—clustering categorical data using summaries**. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 1999:73–83.

89. Andritsos P, Tsaparas P, Miller RJ, Sevcik KC: **LIMBO: Scalable clustering of categorical data**. *Advances in Database Technology-EDBT 2004* 2004:531–532.

90. Tishby N, Pereira FC, Bialek W: **The information bottleneck method**. In *37th Annual Allerton Conference on Communication, Control and Computing*. Urban-Champaign, IL: 1999.

91. Barbara D, Couto J, Li Y: **COOLCAT: an entropy-based algorithm for categorical clustering**. In *In Proceedings of the eleventh international conference on Information and knowledge management*. ACM Press; 2002:582–589.

92. He Z, Xu X, Deng S: **Squeezer: an efficient algorithm for clustering categorical data**. *Journal of Computer Science and Technology* 2002, **17**:611–624.

93. Wang K, Xu C, Liu B: **Clustering transactions using large items**. In *Proceedings of the eighth international conference on Information and knowledge management*. 1999:490.

94. Xiao Y, Dunham M: **Interactive clustering for transaction data**. *Data Warehousing and Knowledge Discovery* 2001:121–130.

95. Yang Y, Guan X, You J: **CLOPE: a fast and effective clustering algorithm for transactional data**. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002:687.

96. Yun CH, Chuang KT, Chen MS: **Adherence clustering: an efficient method for mining market-basket clusters**. *Information Systems* 2006, **31**:170–186.

97. Pedrycz W, Skowron A, Kreinovich V: *Handbook of Granular Computing*. Wiley-Interscience; 2008.

98. Boutros PC, Okey AB: **Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data**. *Brief. Bioinformatics* 2005, **6**:331–343.

99. Beyer K, Goldstein J, Ramakrishnan R, Shaft U: **When Is "Nearest Neighbor" Meaningful?** *IN INT. CONF. ON DATABASE THEORY* 1999:217–235.

100. Fontaine J-F, Priller F, Barbosa-Silva A, Andrade-Navarro MA: **Génie: literature-based gene prioritization at multi genomic scale**. *Nucleic Acids Res* 2011, **39**:W455–W461.

101. Chen Y-A, Tripathi LP, Mizuguchi K: **TargetMine, an Integrated Data Warehouse for Candidate Gene Prioritisation and Target Discovery**. *PLoS ONE* 2011, **6**:e17844.

102. Pers TH, Hansen NT, Lage K, Koefoed P, Dworzynski P, Miller ML, Flint TJ, Mellerup E, Dam H, Andreassen OA, Djurovic S, Melle I, Børglum AD, Werge T, Purcell S, Ferreira MA, Kouskoumvekaki I, Workman CT, Hansen T, Mors O, Brunak S: **Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes**. *Genetic Epidemiology* 2011, **35**:318–332.

103. Watkins NA, Gusnanto A, de Bono B, De S, Miranda-Saavedra D, Hardie DL, Angenent WGJ, Attwood AP, Ellis PD, Erber W, Foad NS, Garner SF, Isacke CM, Jolley J, Koch K, Macaulay IC, Morley SL, Rendon A, Rice KM, Taylor N, Thijssen-Timmer DC, Tijssen MR, van der Schoot CE, Wernisch L, Winzer T, Dudbridge F, Buckley CD, Langford CF, Teichmann S, Göttgens B, Ouwehand WH: **A HaemAtlas: characterizing gene expression in differentiated human blood cells**. *Blood* 2009, **113**:e1–9.

104. Steele BM, Harper MT, Macaulay IC, Morrell CN, Perez-Tamayo A, Foy M, Habas R, Poole AW, Fitzgerald DJ, Maguire PB: **Canonical Wnt signaling negatively regulates platelet function**. *Proc. Natl. Acad. Sci. U.S.A.* 2009, **106**:19836–19841.

105. Iseki H, Takeda A, Andoh T, Takahashi N, Kurochkin IV, Yarmishyn A, Shimada H, Okazaki Y, Koyama I: **Human Arm protein lost in epithelial cancers, on chromosome X 1 (ALEX1) gene is transcriptionally regulated by CREB and Wnt/beta-catenin signaling**. *Cancer Sci.* 2010, **101**:1361–1366.

106. Stalker TJ, Wu J, Morgans A, Traxler EA, Wang L, Chatterjee MS, Lee D, Quertermous T, Hall RA, Hammer DA, Diamond SL, Brass LF: **Endothelial cell specific adhesion molecule (ESAM) localizes to platelet-platelet contacts and regulates thrombus formation in vivo**. *J. Thromb. Haemost.* 2009, **7**:1886–1896.

107. O'Connor MN, Salles II, Cvejic A, Watkins NA, Walker A, Garner SF, Jones CI, Macaulay IC, Steward M, Zwaginga J-J, Bray SL, Dudbridge F, de Bono B, Goodall AH, Deckmyn H, Stemple DL, Ouwehand WH: **Functional genomics in zebrafish permits rapid characterization of novel platelet membrane proteins**. *Blood* 2009, **113**:4754–4762.

108. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data**. *Bioinformatics* 2006, **22**:1122–1129.

109. Ben-Dor A, Chor B, Karp R, Yakhini Z: **Discovering local structure in gene expression data: the order-preserving submatrix problem**. *J. Comput. Biol* 2003, **10**:373–384.

110. Ihmels J, Bergmann S, Barkai N: **Defining transcription modules using large-scale gene expression data**. *Bioinformatics* 2004, **20**:1993 –2003.

111. Tanay A, Sharan R, Shamir R: **Discovering statistically significant biclusters in gene expression data**. *Bioinformatics* 2002, **18 Suppl 1**:S136–144.

List of publications

International journal papers

- Tranchevent L-C*, Capdevila FB*, Nitsch D*, De Moor B, De Causmaecker P, Moreau Y: A guide to web tools to prioritize candidate genes. *Brief Bioinform* 2010:bbq007

- Bonachela-Capdevila F, Tranchevent L-C, Moreau Y, De Causmaecker P: A clustering based preprocessing method for gene prioritization. (submitted)

- Börningen D*, Tranchevent L-C*, Bonachela-Capdevila F*, Devriendt K, de Moor B, De Causmaecker P, Moreau Y: An unbiased evaluation of gene prioritization tools (submitted)

- Bonachela-Capdevila F*, Börningen D*, Tranchevent L-C,* Breckpot J, Brady P, Thienpont B, Laprise C, Deprest J, Devriendt K, Vermeesch JR, de Moor B, Moreau Y, De Causmaecker P: Combination of gene prioritization tools gives an insight into disease gene discovery (submitted)

* equally contributed.

International conference abstracts and oral presentations

- Bonachela-Capdevila F, Tranchevent L-C, Moreau Y, De Causmaecker P: Heuristics for gene prioritization, *Workshop on OR in Computational Biology, Bioinformatics and Medicine*, Prague, Czech Republic, July 8, 2007

- Bonachela-Capdevila F, Broos K, Deckmyn H, De Causmaecker Y: Endeavour and clustering for gene prioritization, *Bloodomics annual meeting*, Leuven, June 23 2008

- Bonachela-Capdevila F, Tranchevent L-C, Moreau Y, De Causmaecker P: Application of clustering techniques in gene priotization using Endeavour, *Mini EURO Conference on Computational Biology, Bioinformatics and Medicine*, Rome, Italy, September 15-17, 2008

- Bonachela-Capdevila F, Tranchevent L-C, Moreau Y, De Causmaecker P: Application of clustering techniques in gene priotization using Endeavour, *EURO XXII*, Bonn Germany, July 5-8, 2009

- Bonachela-Capdevila F, Broos K, Deckmyn H, De Causmaecker Y: Gene Prioritization using Endeavour and Haematlas, *Bloodomics annual meeting,* Cambridge, November 11, 2009

- Tranchevent L-C*, Capdevila FB*, Nitsch D*, De Moor B, De Causmaecker P, Moreau Y: A guide to web tools to prioritize candidate genes A guide to web tools to prioritize candidate genes, *EURO 2010*, Lisbon, July 11-14 2010

- Bonachela-Capdevila F*, Börningen D*, Tranchevent L-C,* Breckpot J, Brady P, Thienpont B, Laprise C, Deprest J, Devriendt K, Vermeesch JR, de Moor B, Moreau Y, De Causmaecker P: Combination of gene prioritization tools reveals new insights into disease gene discovery, *Mini EURO – CBBM*, Nottingham, 13th-15th September 2012

Seminars

- Bonachela-Capdevila F, Tranchevent L-C, Moreau Y, De Causmaecker P: Application of clustering techniques for gene prioritization using Endeavour, *School of Computer Science, Faculty of Sciences, The University of Nottingham*, Nottingham, United Kingdom, 25th May 2009

- Bonachela-Capdevila F, Tranchevent L-C, Moreau Y, De Causmaecker P: Application of clustering in human gene prioritization using CLOPE and Endeavour, *Departement Elektrotechniek*, *KU Leuven*, Leuven, 13th April 2010

- Bonachela-Capdevila F, De Causmaecker P: Clustering and gene prioritization: *Département des sciences fondamentales, Université du Québec à Chicoutimi*, Chicoutimi, Québec, Canada, 9th February 2012

Curriculum Vitae

Francisco Bonachela Capdevila was born in Vic, Barcelona, Spain in 1978. In 2000, he obtained a Master's degree in biochemistry from the University of Granada. In 2003 he obtained a Bachelor's degree in Computer Science from the University of Granada and two years later he completed a Master's program in computer science also in the University of Granada. In 2006 he was a young researcher in the Department of Computing Science and Artificial Intelligence in University of Granada under the supervision of Prof. David Pelta and Prof. José Luis Verdegay and with the support of NiSIS. In the end of 2006, he joined the Combinatorial Optimisation and Decision Support team in KU Leuven Campus Kortrijk under the supervision of Prof. Patrick de Causmaecker and with the support of the Marie Curie Early Stage Training Bioptrain to start a Ph.D. program on bioinformatics.