Gene Prioritization through genomic data fusion

Methods and applications in human genetics

Léon-Charles Tranchevent

Jury:
Prof. dr. A. Bultheel, chairman
Prof. dr. ir. Y. Moreau, promotor
Prof. dr. ir. B. De Moor, co-promotor
Prof. dr. F. Azuaje
(CRP-Santé, Luxembourg)
Prof. dr. J. Vermeesch
Prof. dr. P. De Causmaecker
Prof. dr. ir. H. Blockeel

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

Mai 2011

Legal depot number D/2011/7515/57
ISBN number 978-94-6018-355-3

Preface

This thesis summarizes my research work as a PhD student in the bioinformatics group, from the SCD research division at the Electrical Engineering Department ESAT, of the Katholieke Universiteit Leuven. During these years, I had the opportunity to learn from internationally renowned researchers, and to collaborate with sensible scientists in a very stimulating yet relaxing environment. I will remember my stay in Leuven as a memorable experience and I am taking this opportunity to thank all the people I have met and worked with.

My deepest gratitude goes to my promotors Prof. Yves Moreau and Prof. Bart de Moor, for giving me the opportunity to join the bioinformatics group to start a PhD under their supervision. In the early days, Yves helped me to find my way through the (sometimes tortuous) labyrinth of bioinformatics research. Enthusiastic discussions with him are an eternal source of new research ideas, and refinement of existing projects. His expertise was very much appreciated and he inspired me in many ways (from mathematics and machine learning to biology and human genetics). I would also like to express my gratitude to Bart for his support along my doctorate. I acknowledge in particular his help regarding all the administrative and funding work. I would also like to recognize the influence of Prof. Hendrik Blockeel on my research, his master course gave me more insights into machine learning methods and it remains a source of inspiration for my machine learning oriented work. I am grateful for his supervision of my doctoral research from the very beginning and also for joining my examination committee. May I extend my sincere thanks to Prof. Joris Vermeesch. He made possible effective collaborations between our bioinformatics group and his laboratory for cytogenetics and genome research. The BeSHG course he organized has given me a clearer picture of human genetics problematics; the introduction of the present thesis is mainly based on that course. I thank him for joining the examination committee and for reading the manuscript with a human genetics point of view. I am particularly honored that Prof. Francisco Azuaje has accepted to join my examination committee. His research is oriented towards the development of 'systems biology' approaches, with direct applications to genetics and medical problematics, which is also the focus of my research. Several key discussion points were added to the present thesis

upon his suggestions. I would also like to express my gratitude to Prof. Patrick De Causmaecker for having accepted to join my examination committee. I have particularly appreciated his thorough review of my manuscript and the machine learning perspective he brought. His suggestions, I believe, made the manuscript clearer and better structured.

I am not forgetting Prof. Enrico Carlon, who gave me the opportunity to join his group for my master internship in Lille. In his team, I discovered how interesting, exciting and challenging bioinformatics research can be. It was very nice to work in an interdisciplinary environment, at the borders of computer science, mathematics, physics, and biology. I am also grateful for his help to prepare my PhD presentation and interview, his precious advices helped me to convince Yves to hire me. I thank Prof. Adhemar Bultheel for being the chairman of the examination committee, and Prof. Robert Vlietinck for joining my supervisory committee. I would also like to acknowledge Ida Tassens, Ilse Pardon, Mimi Deprez, John Vos, Veronique Cortens, Eliane Kempenaars, Evelyn Dehertoghe, and Lut Vander Bracht for their support regarding all the administrative tasks, it was not always a long and quiet river and their help was really appreciated.

Five years were enough to create and exploit many fruitful collaborations with people from different laboratories and with different backgrounds. In particular, I am indebted to Prof. Stein Aerts, Bert Coessens, and Peter Van Loo for their initial support. They have taken the time to answer the many questions I had, allowing me to efficiently take over their gene prioritization work. The kernel based research was done in close collaboration with Prof. Tijl de Bie, and Shi Yu. I am particularly honored I could collaborate with Shi Yu, a tireless researcher who masteries kernel based methods. He significantly influenced my current gene prioritization work as well as the research I want to perform in the near future. A special thank goes to Alexander Griekspoor and Prof. Dietrich Rebholz-Schuhmann, who kindly accepted me in their text mining group at the EBI. It was a short but fruitful stay, working with renowned researchers in a peaceful environment.

Another important collaborator is Lieven Thorrez, who is always able to define challenging biological questions that can be answered using a combination of computational and wet lab methods. I am thankful for the many projects he shared, including the master thesis of Hui Ju Chang, and the projects with Katrijn Van Deun. I take this opportunity to also thank Roland Barriot, Steven Van Vooren, Sonia Leach, Francisco Bonachela Capdevila, Daniela Nitsch, Sylvain Brohée, Joana Gonçalves, Xinhai Liu, Ernesto Iaccucci for keeping me busy with prioritization related work. They all contributed to the work described in the present thesis, and I feel lucky I had the opportunity to collaborate with so many great people. In particular, I acknowledge Sonia and Roland who kindly shared their valuable PhD experience with me. For the applications to real biological questions, I could rely on collaborations with Bernard Thienpont, Jeroen Breckpot, Irina Balikova, Paul Brady, Julio Finalet, Beata Nowakowska, Lieven Thorrez,

Prof. Hilde Peeters, Prof. Mathijs Voorhoeve, Prof. Koen Devriendt, Prof. Stein Aerts, Prof. Bassem Hassan, and Prof. Frans Schuit. I thank them for taking the time to understand the computational solutions we proposed, and for having the patience to explain (or re-explain) the 'basics' of genetics. I sometimes say that a computational method is rather useless if not applied to real problems, I sincerely appreciate that they biologically validate our methods. I would also like to thank Sven Schuierer, Uwe Dengler, Wim De Clercq, Berenice Wulbrecht, Domantas Motiejunas, Koen Bruynseels, for the collaborations with industrial partners. I also thank our system administrators, in particular Edwin Walsh and Maarten Truyens, for their support regarding the IT structure behind our tools. They both assisted me in the challenging task of keeping the tools up and running at all time.

It was a pleasure to work in the bioinformatics group all these years, I shall remember Thomas Dhollander, Karen Lemmens, Joke Allemeersch, Peter Monsieurs, Wouter Van Delm, Olivier Gevaert, Tim Van den Bulcke, Ruth Van Hellemont, Raf Van de Plas, Liesbeth van Oeffelen, Leander Schietgat, Fabian Ojeda, Toni Barjas-Blanco, Peter Konings, Jiqiu Cheng, Tunde Adefioye, Nico Verbeeck, Alejandro Siffrim, Arnaud Installe, Dusan Popovic, Minta Thomas, and Yousef El Aalamat. I also have a thought for the Bioptrain group, Daniel Soria, Pawel Widera, Enrico Glaab, Andrea Sackmann, Marc Vincent, Matthieu Labbé, Linda Fiaschi, Jain Pooja, Aleksandra Swiercz, and Prof. Jon Garibaldi. It was really nice to meet in exotic places, and to share about our research experience.

Ultimately, I thank my beloved family and my friends for their support and encouragement during the course of my doctoral research. In particular, I thank my parents for giving me the opportunity to achieve my dreams. Although words fail me to express my appreciation, I dedicate this thesis to Valerie, who keeps me going with her support, patience and love.

Léon-Charles Tranchevent
Leuven, May 2011

Abstract

Unravelling the molecular basis underlying genetic disorders is crucial in order to develop effective treatments to tackle these diseases. For many years, scientists have explored which genetic factors were associated with several human traits and diseases. After the completion of the human genome project, several high-throughput technologies have been designed and widely used, therefore producing large amounts of genomic data. At the same time, computational tools have been developed and used in conjunction with wet-lab tools to analyze this data in order to enrich our knowledge of genetics and biology.

The main focus of this thesis is gene prioritization, that can be defined as the identification of the most promising genes among a list of candidate genes with respect to a biological process of interest. It is a problem for which large quantities of data have to be manipulated, which typically means that it has to be done *in silico*. This thesis describes two gene prioritization methods from their theoretical development to their applications to real biological questions.

The first part of this thesis describes the development of two data fusion algorithms for gene prioritization respectively based on order statistics and kernel methods. These algorithms have been developed for human and also for reference organisms. Ultimately, a cross-species version of these algorithms have been developed and implemented. Integrating genomic data among closely related organisms is relevant since many researchers are studying human indirectly through the study of reference organisms such as mouse or rat, and are therefore producing mouse/rat specific data, that is still relevant in human biology. Our method can integrate more than 20 distinct genomic data sources for five organisms and is therefore one of the first cross-species gene prioritization method of that scale.

Only a fragment of all the computational tools developed each year specifically for biology are still maintained after three years, and even less are used by independent researchers. The second part of this thesis focuses on the benchmarks of the proposed methods, the development of the corresponding web based softwares, and on their application to real biological questions. By making our methods publicly available, we make sure that interested users can apply them for their

own problems. In addition, benchmarking is needed to prove that the approach is theoretically valid and can estimate how accurate are the predictions. Ultimately, the inclusion of our computational method within wet-lab workflows show the real usefulness of the approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This introductory chapter presents several basic concept of human genetics and bioinformatics that are at the core of the work described in the present dissertation. Section 1 points out some current challenges of human genetics, and describe how bioinformatics methods are used today in conjunction with wet lab methods to fasten the research process. Section 2 presents more in details the gene prioritization problem that is the focus of this thesis and describes the challenges and objectives of that field. Section 3 describes the genomic data sources that are at the core of the gene prioritization problem. Section 4 summarizes the options available as for benchmark and validation of the algorithms. Section 5 outlines the content of the next chapters.

1.1 Human genetics

1.1.1 Molecular genetics

The basic unit of the human body is the cell, an adult human body contains billions of cells. Almost every cell contains in its nucleus the human genome, physically a set of 23 pairs of chromosomes. Chromosomes are very long condensed stretches of deoxyribose nucleic acid (DNA). Beside its sugar and phosphate backbones, DNA is made up of four distinct nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). In total, the 23 human chromosomes represent 3 billions nucleotides. Genes are chromosomal fragments that contain all the necessary information to create proteins that are the real workers of the cells. According to the latest estimation, the human genome contains between 20000 and 25000 protein coding genes [55]. Proteins are acting either alone or within complexes to achieve precise

functions inside and outside the cells. An example is the AMY1A gene, located on chromosome 1 and associated to a protein termed amylase, an enzyme that digests starchy food. Although AMY1A is present in every human cell, it is mostly active in the salivary gland and amylase is therefore mostly present in the saliva. Most human cells have 46 chromosomes organized in 23 pairs, meaning that each gene is usually present in two copies (one on each chromosome), exception made of the genes located on the sexual chromosomes X and Y. The raw information of a gene resides in its coding sequence, that is the nucleotide sequence that encodes for the gene products themselves (*e.g.*, proteins), there are however extra elements that control when and where genes are expressed (including other non protein coding genes).

In theory, every chromosome is present in two copies in every cell, and therefore the genes are also present in two copies. However, in practice, it has been noticed that several genetic alterations can occur:

- Copy Number Variants (CNVs) or Copy Number Changes (CNCs): a chromosomal region can be deleted (*i.e.*, a single copy is present), double deleted (*i.e.*, no copy at all), duplicated (*i.e.*, a third copy is observed), or amplified (*i.e.*, at least two supplementary copies are present). When the region expands over an entire chromosome, the terms used are triploidy (three copies), tetraploidy (four copies) and haploidy (single copy).
- Structural rearrangements: this refers to the reorganization of the sequence, the overall content stays the same (no deletion, no duplication) but the order is changed. For instance, a chromosomal region can be translocated, meaning that the region is removed from its original location and inserted into another location, possibly on a different chromosome, and possibly disrupting the sequence of a gene.
- Single Nucleotide Polymorphism (SNP): a single nucleotide can be altered (*i.e.*, mutated, deleted or inserted), changing the gene sequence possibly at a key position, altering therefore its function.
- Epigenetic modification: epigenetic refers to all factors that affect the use of the genes without affecting their raw DNA sequences. For instance, genes can be modified through DNA methylation (addition of a methyl group to cytosine nucleotides) or chromatin modification (*e.g.*, via histone modification).

Most of the difference observed between human individuals at the genome level (mostly SNPs and CNVs) are accountable for the differences observed at the phenotypic level (*e.g.*, eye and hair color, blood type, height). These genomic variations are frequently observed and are not linked to diseases as shown by Redon *et al.* [200]. An example is a locus on chromosome 1 that contains the AMY1A gene, responsible for the production of amylase, the saliva enzyme that

digests starchy food. Perry *et al.* have studied this region in seven populations and noticed that it is usually repeated several times [189]. Moreover, they show that the european-american and japanese individuals would have, on average, more copies than individuals from the Yakut and Biaka populations. The maximum is observed for the japanese population with up to 16 copies instead of the expected two copies. The authors also show that this could be explained by their very different diet (historically a lot of starchy food for the europeans and japanese populations and a low starchy diet for the Yakut and Biaka populations). There are however genomic alterations that can predispose or even cause diseases, these are the focus of medical genetics.

1.1.2 Medical genetics

Human genetics refers to the study of biological mechanisms in order to explain the similarities and the differences among human beings. Medical genetics refers to the application of human genetics to medicine. That is how biological processes relate to human diseases. A disease can be defined as a set of observable characteristics (termed traits or phenotypes), and is said to be genetic when one contributing factor is genetic, that is when the phenotypes can be associated to the genome of the patients and more precisely to the genomic alterations that can be observed. Beside genetic factors, environmental factors such as smoking and diet can also contribute.

An example of genetic disease is cystic fibrosis, that mainly affects the lungs and the digestive system. The malfunction is due to the abnormal accumulation of a thick mucus that prevents the organs to achieve their function properly. This condition has been linked to a locus on chromosome 7 where lies the CFTR gene (Cystic Fibrosis Transmembrane conductance Regulator gene) [203, 256]. More precisely, it has been observed that any individual with two abnormal copies (*e.g.*, with a mutation) of the gene has cystic fibrosis.

Medical geneticists aim at unravelling the molecular basis underlying genetic disorders, in order to understand what is exactly happening down to the molecular level. A better understanding of the disease players and their mode of action is however only the first step towards the development of effective treatments to tackle these diseases. Although the genetic defects that underlie a disease are virtually present in every single cell of a patient, it is still possible to develop effective treatments that will eliminate or reduce the effects of the disease. The treatments can possibly intervene at the gene level (*e.g.*, to replace a mutant allele), at the mRNA level (*e.g.*, to keep the expression of a mutant RNA under control), at the protein level (*e.g.*, to replace the defective protein) or even at the clinical level (*e.g.*, surgery or transfusion). At the protein level, an example is the administration of insulin to type 1 diabetes mellitus patients to remedy to the destruction of

the beta cells of the pancreas that are producing insulin. Another example is phenylketonuria for which a dedicated diet combined to a light medication can treat the disease with almost no side-effects [135, 137].

We usually refer to the beginning of genetic with the work of Mendel, a monk who studied the heredity of physical traits in peas during the 19th century. Medical genetics, however, had its start at the very beginning of the 20th century, with the recognition that Mendel's laws of inheritance explain the recurrence of genetic disorders within families [56, 60, 247] and therefore the recognition of the hereditary nature of several human diseases. An example is hemophilia, a disorder that impairs blood coagulation, that was already reported in antiquity and for which the underlying factors remained unknown for centuries. The discovery of the first factor (factor V), in 1947 [176], and the subsequent discoveries of additional factors proved the hereditary nature of hemophilia.

In the second half of the 20th century, many studies have been performed in order to discover which genomic alterations are responsible for which disorders mainly through the study of syndromes such as Marfan [15, 37, 44, 155] and Ehlers-Danlos syndromes [168, 254]. At that time, the techniques used, such as Southern blot [223] and Polymerase Chain Reaction (PCR) [125], were mostly wet lab based and computer science had little if no role to play in this analysis. In 1966, Victor McKusick created the Mendelian Inheritance in Man (MIM), an extensive catalog of human genes related to genetic disorders that quickly became a reference in genetics [156]. As of today, the online version of this catalog, the Online Mendelian Inheritance in Man (OMIM) represents a comprehensive catalog of the current knowledge in medical genetics (more than 13000 genes and 4000 phenotypes) [94, 95, 157].

A major breakthrough in genetics was the sequencing of the human genome (first draft in 2001 [131] and its completion in 2003 [55]). This task revealed the three billions nucleotides that encode our genome, and the 20000 to 25000 genes that make us human. However, rather than the end of genetics, this was more the beginning of a new era, the post-sequencing era. Indeed, the knowledge of the genome sequence has led to the development of high-throughput technologies such as micro-arrays that measure the expression level of thousands of genes concurrently. The use of these technologies has considerably increased the amount of genomic data available meaning that the main task is today to harvest the fruits that are hidden in this data. Altogether, this means that computational biology is now playing an important role and this thesis serves as an illustration. The computational approach developed in this thesis has been integrated into wet lab based workflow (chapters 3, 4, and 8).

Moreover the focus of the studies has shifted towards a 'systems biology' approach. Until recently, reductionist approaches were often used in biology to break down a complex system into simpler components that were then analyzed individually.

This very successful approach is now complemented by integrated approaches that can analyze a complex system at once by taking advantage of the genome wide data produced in the post-sequencing era and of elaborated computational tools that can deal with its complexity. Nowadays, 'systems biology' is becoming the standard approach in computational biology and this thesis also illustrates this by integrating several data sources in order to unravel the biological mechanisms at the disease level.

Computational biology

As stated in the previous section, a perfect understanding of the molecular mechanisms that underlie a genetic disorder is crucial in order to develop efficient treatments. This knowledge about the molecular and cellular processes is nowadays increasing fast due to the use of systems biology based approaches [20, 21, 48, 38, 150]. One of the main objectives is to define efficient algorithms that combine the existing knowledge with raw data in order to create novel hypothesis to be experimentally assayed and eventually enrich our knowledge. Therefore these algorithms are fully integrated within a workflow that merges together wet lab tasks and computational tasks. Such processes are cyclic so that the enriched knowledge can be used to create additional hypotheses that will undergo the same validation. The cycle presented in figure 1.1 represents a typical computational biology approach that mixes together wet lab work with *in silico* methods.

In the recent years, several computational tools that target biologists and human geneticists have been developed, this includes tools to organize and query the scientific literature (Pubmed, GoPubmed [66]), or expression data repositories (Gene Expression Omnibus [70, 27], ArrayExpress [181]) knowledge bases (Ingenuity® and MetaCore™) or collaborative knowledge bases (CHDWiki [28], see also chapter 8, WikiGenes [101]), tools to analyze and interpret high-throughput data such as expression data (GeneSpring, ArrayAssist®, R / Bioconductor [81]), or tools with multiple functionalities among the ones cited (DECIPHER [76]).

An example of computational tools developed for human genetics is Bench™, Cartagenia's platform for Array Comparative Genomic Hybridization (array CGH). It is made of two components. First, an intelligent repository that allows users to manage and visualize results from various genetic screening assays, from array CGH data to next generation sequencing platforms. Second, a software solution that help users to rapidly interpret the copy number alterations in patient samples, and to assess their clinical relevance and impact in patient and population genotypes.

Another example is DECIPHER, the DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources. DECIPHER collects clinical information about chromosomal microdeletions/duplications/insertions, translocations and inversions and displays this information on the human

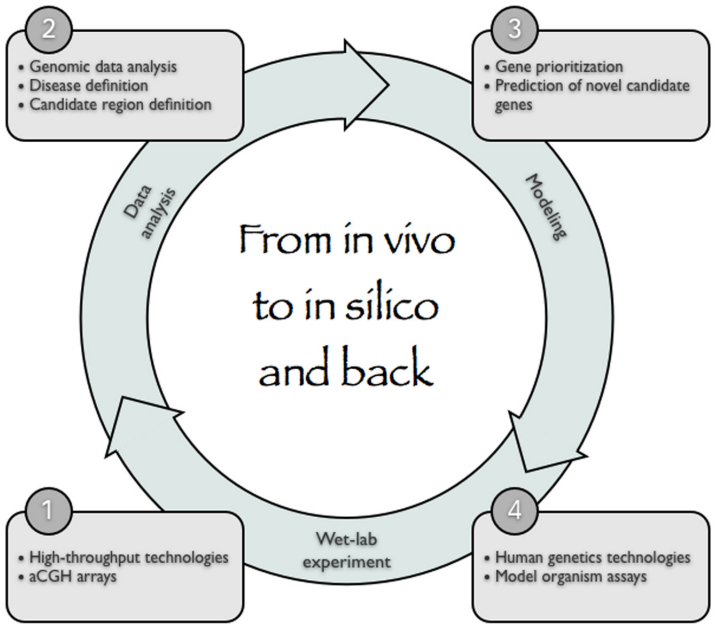

Figure 1.1: The research cycle that involves wet lab experiments and computational biology. Computational tools are used to analyze the data and to produce novel hypothesis. Wet lab experiments are used to produce data and to validate the hypothesis. The boxes describe such a workflow that involves gene prioritization. (1) In the first step, high-throughput technologies are used to produce genomic data that is further used by the gene prioritization approach. In addition, the array CGH technology can be used to define a region to investigate. (2) In a second step, the genomic data produced is analyzed and organized, and the biological hypothesis is defined as a computational problem. (3) In the third step, gene prioritization is used to predict novel candidate genes. (4) The predictions are experimentally validated using sequencing or model organism knock-outs. The analysis of this data is then enriching the current knowledge.

genome map with the aim of improving medical care and genetic advice for individuals/families with submicroscopic chromosomal imbalance and facilitating research into the study of genes which affect human development and health. DECIPHER is a consortium that gathers several research groups and hospitals world wide, meaning that the information is shared among the members to fasten the research process.

Another example is the gene prioritization problem and is introduced in details in the next section.

1.2 Gene prioritization

Gene prioritization has been defined as the identification of the most promising genes among a list of candidate genes with respect to a biological process of interest. It has been designed to augment the traditional disease gene hunting techniques such as positional cloning. The motivation behind gene prioritization is that, very often, the gene lists that are generated contain dozens or hundreds of genes among which only one or a few are of primary interest. The overall objective is to identify these genes, however the experimental validation of every candidate individually is expensive and time consuming, and it is therefore preferable to define, in a preliminary step, the most promising candidate genes and, in a second step, to experimentally validate these genes only. This conceptual approach is illustrated in figure 1.2.

The concept of gene prioritization was first introduced in 2002 by Perez-Iratxeta *et al.* who already described the first computational approach to tackle this problem [185]. Since then, many different computational methods that use different strategies, algorithms and data sources, have been developed [275, 2, 110, 4, 50, 273, 211, 268, 205, 152, 51, 240, 102, 248, 272, 82, 265, 126, 196, 80, 163, 148, 39, 239, 77, 235, 169, 236, 187, 186]. Some of these approaches have been implemented into publicly available softwares allowing their use by researchers worldwide. Eventually, several of these approaches have been experimentally validated including the approaches presented in this dissertation. A thorough review of the publicly available gene prioritization web tools is presented in chapter 2.

1.2.1 Context

This section presents the motivation behind the work presented in this thesis with the description of three research or clinical practice situations in which there is a need for gene prioritization. There are of course many other possible applications, some of them are described in chapter 4, 8, and 9.

Chromosomal aberration in a patient with a genetic condition

In clinical practice, geneticists are often investigating a cohort of patients who share a genetic condition and for which a recurrent chromosomal aberration has been detected through the use of array CGH. The aim is then to discover which genes are responsible for the observed phenotype and, therefore, to get a better understanding of this phenotype. The chromosomal region corresponding to the aberration often contains dozens of genes among which only one or a few are believed to be responsible for the genetic condition under study. Typically, the validation of individual genes can occur through sequencing in a distinct cohort of

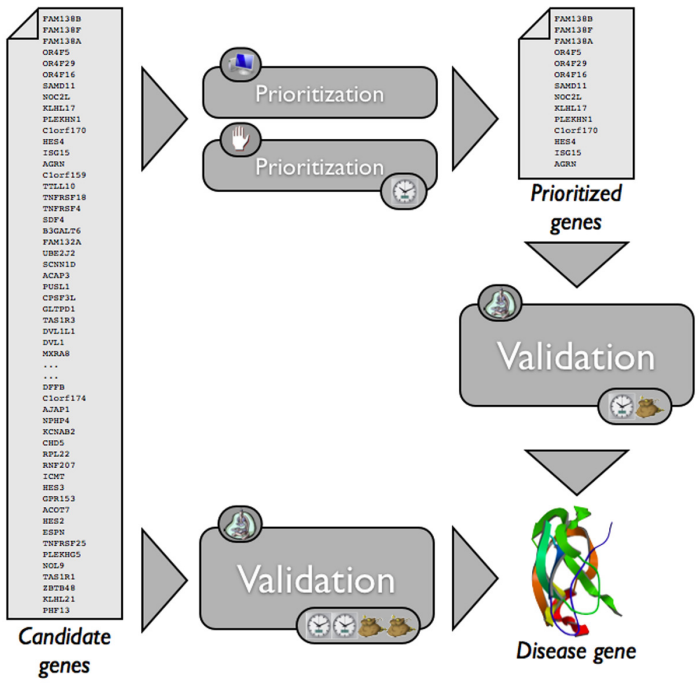

Figure 1.2: The concept of gene prioritization. The starting point is a large list of candidate genes (on the left) among which only one or a few are really of primary interest with respect to the biological process of interest (*e.g.*, a genetic disorder). The goal is to identify this gene (bottom right corner). One solution is to experimentally validate all the candidate genes but this can be very expensive and time consuming (bottom workflow). Another solution is to prioritize the candidate genes using a computational approach at almost no cost, and in a second step, to validate only the most promising genes (top workflow). The second strategy has the advantage of being cheaper and less time consuming.The prioritization can be achieved manually or automatically through the use of dedicated computational programs. The latter solution is even faster.

patients (who do not exhibit the aberration) or through the use of model organism based experiments (*e.g.*, knock out). Although these bio-technologies are getting cheaper and cheaper, it is still expensive for most labs to perform this validation for dozens of genes at the same time. In that case, candidate gene prioritization can be performed beforehand on the chromosomal region to determine the most promising genes to validate. Only the most promising candidate genes will then be experimentally assayed. An example is the prioritization of an atypical DiGeorge syndrome region on chromosome 22q11 that encompasses 68 genes, followed by the

validation of the most promising candidate genes through knock out in zebrafish embryos [4] that leads to the identification of YPEL1 as a putative novel DGS gene (see also chapters 3 and 8).s

Differential expression of genes in a disease tissue

It is sometimes not possible to restrict the analysis to a particular chromosomal region and the solution might then be to consider the whole genome. An efficient way to discover new disease genes genome wide is to compare the gene expression levels between a diseased tissue and a reference tissue. There is a plethora of methods to detect differential expression such as fold change, t-test, SAM and Cyber T. These methods have been extensively compared [178, 242, 166] but in most cases, large lists of differentially expressed genes that contain hundreds of genes are generated. Similarly to the first situation, only a few differentially expressed genes are directly involved in the disease under study, and the other genes are the results of perturbations happening more downstream of the regulatory cascade. It is again expensive to validate hundreds of genes and prioritization is therefore key. An example of such gene expression study is shown in Aerts *et al.* [4] (see also chapter 3).

Linkage analysis

Identifying novel disease genes genome wide can also be achieved through positional cloning strategies. Traditional positional cloning strategies involve first a linkage analysis followed by a closer investigation of the genes located in the region that is linked to the disease of interest. A linkage analysis is the study of genetic markers in a population and their correlation with a disease of interest. The markers that do exhibit correlation with the disease (*i.e.*, low recombination) indicate the presence of a disease causing gene in the neighborhood. Typically, a region of a few to several millions of bases around the marker is considered to harbor the disease gene. The problem is then to find the disease causing gene among the candidate genes and again gene prioritization can be performed. Linkage studies are very popular and have allowed a number of important discoveries, for instance for multiple sclerosis [92, 93], insulin-dependent diabetes mellitus [88, 96] and various X linked disorders [24, 26], they are now complemented by array CGH in clinical routines.

1.2.2 Algorithms

The traditional approach for gene prioritization is to perform a manual search of what is known about the candidate genes and to manually select the ones that seem more interesting based (i) on the small amount of data available at that time

and (ii) on the expertise of the user. The main problem of this approach is related to the amount of genomic data available nowadays in the post-sequencing era. More and more organisms have seen their genome sequenced and, more important, annotated. Many high-throughput technologies such as micro arrays [209, 132] have been developed and widely used to screen the expression level of hundreds of different conditions genome wide. This is in contrast with the pre-sequencing era when only little information was available about each gene. This makes the manual analysis described above at most painful, if not impossible at all. To circumvent that problem, the development of *in silico* gene prioritization solutions has received a lot of interest from the bioinformatics community in the last decade. Most of the gene prioritization methods are based on the automation of the traditional approach. At the heart of these methods is the 'guilt-by-association' concept: the most promising candidate genes are the genes that are similar to the genes already known to be linked to the biological process of interest [219, 87, 115].

This 'guilt-by-association' concept has already been used in the past to align gene sequences. Before any genome was sequenced, small DNA sections were investigated individually to assess their function. It was soon discovered that the function is directly linked to the DNA sequence content and that it is therefore possible to predict the function of an unknown sequence by looking at its similarities to sequences with known functions. This approach is implemented in the Basic Local Alignment Search Tool (BLAST) in 1990 [12]. A gene prioritization strategy can be seen as an extension of the Blast approach [12] in which predictions are made by looking at the similarities between DNA or protein sequences. For example, when studying type 2 diabetes (T2D), KCNJ5 appears as a good candidate through its potassium channel activity [111], an important pathway for diabetes [252], and because it is known to interact with ADRB2 [133], a key player in diabetes and obesity. This notion of similarity is not restricted to pathway or interaction data but can rather be extended to any kind of genomic data. Although the early gene prioritization methods relied on a single or a few data sources, nowadays most of the gene prioritization methods take advantage of several data sources. It is therefore of crucial importance to define an elegant data fusion strategy.

'Integrative genomics' or 'integromics' is the area of research that focuses on data integration [253]. It became very popular after the first high-throughput technologies started to produce a huge amount of data. The motivations behind data integration are multiple.

1. The first one is linked to the missing data problem, the combination of several data sources with missing data is likely to increase the overall coverage therefore reducing the genes with missing data.

2. Second, a synergetic effect is expected: the whole can be more than the sum of its components, meaning that the combination of several data sets can perform better than using any of the data set alone. A question that arises

however is the number of data sources to combine in order to reach critical power, the rule might not be to inlude as many data sources as possible stated by Lu *et al.* who found that 4 out of 16 features is optimal for PPI prediction [144]. This issue is further discussed in chapter 5.

3. Third, different data sources may be contradictory, by integrating them, a consensus can be found thus (i) favoring the predictions that are backed up by multiple data sources (*i.e.*, giving strong confidence) and (ii) rolling out the spurious predictions that are present in only one data source (*i.e.*, assimilated to noise).

4. Fourth, with data integration, an overall strong prediction score can be obtained through the combination of several weaker prediction scores, which the study of a single data source alone would not allow.

Nowadays the term 'integromics' is not used anymore since almost all 'systems biology' approaches are integrating multiple data sources [108, 63, 136, 9]. However, the key challenges remain, they are the integration of different data types using different formats [243], the data quality control, possibly involving correlation analysis, and the design of a dedicated algorithm (no 'one size fits all' paradigm). One important aspect of data integration is that it should not introduce a bias towards well studied genes, meaning that even the poorly characterized genes can be highly prioritized. Another aspect is the use of algorithms that are assuming independence between the data sources while the underlying data sources are usually correlated. These weak correlations can bias the results through a rumor propagation like system therefore increasing the number of false positives. A third important aspect in data integration approach development is the validation, either *in silico* or through wet lab experiments. There exist multiple algorithms to perform data integration including voting system [249, 90], naive Bayesian integration [113, 47, 246, 238], likelihood-based algorithms [208], decision trees [260], and support vector machine (SVM) [130, 31]. For example, Troyanskaya *et al.* have developed MAGIC (Multisource Association of Genes by Integration of Clusters), a general framework that uses formal Bayesian reasoning to integrate heterogeneous types of high-throughput biological data for gene function prediction. To build the network, they use yeast protein-protein interactions from GRID, pairs of genes that have experimentally determined binding sites for the same transcription factor (from the Promoter Database of Saccharomyces Cerevisiae - SCPD), and gene expression data (analyzed through clustering). The inputs of the system are gene clusters based on co-expression, co-regulation, or interaction. The Bayesian network then combines evidence from input clusters and generates a posterior belief estimating whether each gene i-gene j pair has a functional relationship. The present thesis discusses two algorithms: data fusion via Order Statistics (OS) and support vector machine (SVM).

Data integration can be realized at different levels. This section and figure 1.3 describes three integration schemes. In the first option the integration happens at the raw data level, it is then an 'early integration' or 'full integration' scheme in which the data sources are combined before applying any algorithm (*e.g.*, modeling / training) in order to create a single input data source. An example is the merging of several small-scale protein-protein interaction datasets into a global larger dataset. This scheme has the advantage of being rather easy to implement when the underlying data structure allows such integration but it is not always the case. It is sometimes preferable to perform the data integration within the algorithm itself, this is termed 'intermediate integration' or 'partial integration'. Dedicated algorithms such as kernel based SVM integrate several data sources during the learning process. Then they produce a single outcome based on all (or a subset of) the inputs. The last option is integration at the knowledge level, it is then a 'late integration' or 'decision integration' scheme. In this case, the algorithm is applied individually to each data source. It is only then that the algorithm outcomes (*e.g.*,hypothesis, predictions, decisions) are combined to generate a global outcome.

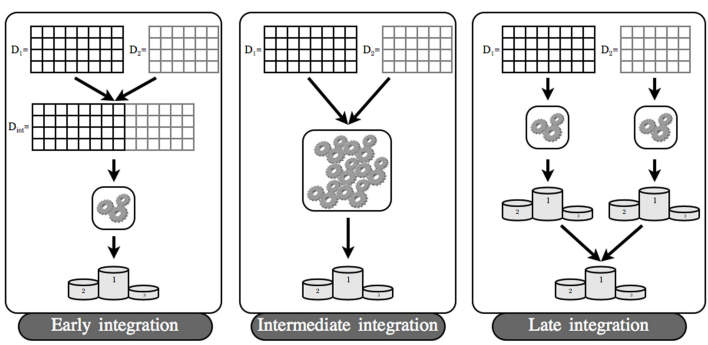

Figure 1.3: Data integration schemes. (Left panel) Early integration. The integration happens at the raw data level before applying the algorithm on the merged data source to produce a single outcome. (Middle panel) Intermediate integration. The integration is realized within the algorithm that accepts several data sources as input and produces a single outcome. (Right panel) Late integration. The algorithm is applied to the data sources independently. The outcomes are then integrated to create a global outcome.

Several candidate gene prioritization have been defined in the last decade, and they can be divided in three main categories: *ab initio* methods, classification methods, and novelty detection methods.

Ab initio methods

The concept of *ab initio* methods is to select candidate genes based on a set of properties that are defined *a priori* to correspond to the disease under study. These properties are often based on physical features (*e.g.*, chromosomal location, gene length) and on expression data (*e.g.*, positive or negative expression in a tissue of interest). After the selection, only the genes that satisfy all the properties are considered as promising candidate genes. The use of several properties in conjunction allows a more conservative filtering that retains only the best candidate genes. The main limitation of these methods is that filters act as binary classifiers and do not allow fine candidate prioritization. For example, it is sometimes difficult to define the optimal properties that limit the number of false positive genes (non interesting genes included) and false negative genes (interesting genes rejected) when using gene expression data that is often noisy. A second limitation is that all the candidate genes that satisfy the properties are all equal and there is no way to estimate which genes should be experimentally validated first.

An example is the study of Parkinson's disease by Hauser *et al.* who used two filters to identify novel candidate genes. The first filter was based on Serial Analysis of Gene Expression (SAGE) data to identify the genes that are expressed in substantia nigra and adjacent midbrain tissue. The second filter identifies the genes that lay within five large genomic regions identified through linkage analysis. These two filters are then combined to identify 402 promising candidate genes for Parkinson's disease [97]. Franke *et al.* created additional filters based on functional data (from Gene Ontology [17]) to select the functionally related genes and association based data to select the genes that are associated to the disease in sub-populations. They have implemented their method into a publicly available software termed TEAM and have applied it to celiac disease and were able to select 120 candidate genes [77]. More recently, Bush *et al.* have developed Biofilter, that integrates even more databases that contain pathway annotations (*e.g.*, KEGG [118, 120, 119]) and protein-protein interactions (*e.g.*, DIP [263, 262, 264, 206]) [43].

Classification methods

Classification starts with a training step, in which the classifier is trained with gene sets that correspond to the distinct classes. In a second step, the candidate genes (unlabeled) are distributed into the classes according to their properties. For gene prioritization, most methods use binary classification, and the two classes correspond to the positive genes (known to be involved in the process under study) and the negative genes (known not to be involved in the same process). The main challenge of these methods resides in the assembly of the negative training set. It is often very difficult to guarantee that a gene is not involved in a biological process, our knowledge is often not elaborated enough to backup such statements [46]. Some

studies have proposed to use unrelated diseases to built the negative training set but that could potentially induce spectrum bias in the classification (negative genes selected not representative of the whole negative population). Others techniques have been developed to tackle that problem including the use of randomly selected genes together with repetitions of the classification process (*e.g.*, for a genome wide approach, use one third of the genome as negative genes to classify the remaining two thirds, and repeat the procedure for the other two thirds [162]). However, this problem can sometimes be awkward given that some classification methods are not efficient with unbalanced data (which we have in our case). A related issue is that, in practice, the number of known genes for one disease is often too small to constitute a reliable positive training set.

A proposed solution is to use a group of closely related diseases (*e.g.*, all cancers [226, 183, 277], dominant versus recessive inheritance [46]) or even to use all diseases at once [3, 142]. Several classification methods also associate a score with every candidate gene that makes the method more suitable for prioritization [3, 142]. For instance, Adie *et al.* have used sequence based features (*e.g.*, gene length, UTR lengths, number of exons, CG content, homology, CpG islands) and a decision tree to classify the human genes between likely disease genes and unlikely disease genes. They train using all disease genes together and show in addition that smaller training sets can not be used efficiently so that analysis are restricted to large group of diseases such as oligogenic or monogenic disorders. Before that, Lòpez-Bigas *et al.* have used protein sequence features (*i.e.*, protein length, phylogenetic extent, degree of conservation, and paralogy) and a decision tree again to reach the same goal. Training was performed using all disease genes from OMIM but the authors do not report any experiments with disease specific training sets.

Novelty detection methods

Novelty detection methods are a variant of classification methods for which no negative training set is needed. Oppositely, they only rely on the positive training set. Candidate genes are then ranked according to their similarities to the training genes. This positive set most often consists of genes that are known to be involved in the disease under study, but it can also be derived from a set of keywords that describe precisely the genetic condition of interest. In the latter case, the candidate genes are ranked according to their similarities to the keywords mainly through text mining. This category is the one that has received the most of interest in the last decade and several strategies have been defined mainly following the early classification based methods [4, 205, 240, 80, 163, 148, 239, 187, 186]. The main characteristic of novelty detection methods is that they are less conservative since they usually rank the genes instead of filtering them, as opposed to *ab initio* methods.

For instance, Turner *et al.* have developed POCUS, a tool that prioritizes candidate genes based on their InterPro domains and Gene Ontology terms that are shared with the genes from the positive training set (no negative training set needed). This method allows candidate gene prioritization using disease specific training sets and therefore was benchmarked with 29 OMIM diseases. Rossi *et al.* developed TOM, a tool that uses expression data and functional annotations together to predict the most interesting candidate genes with respect to a biological process. This process is defined by a set of genes known to play a role in it, no negative set has to be defined. The work presented in this dissertation mostly focuses on novelty detection methods.

Related strategies

A first category of related strategies contains the microarray analysis tools (*e.g.*, GeneSpring, ArrayAssist®, ArrayStar, Mapix, Qlucore Omics Explorer, Axon GenePix, and PathwayArchitect‡). These tools allow users to analyze large list of genes. There are however several differences:

1. They are not making use of various genomic data sources and usually rely on expression data alone (or in combination with phenotypic data). Gene prioritization aims at combining many data sources, including, for instance, literature data, functional annotations, sequences and regulatory information.
2. Microarray analysis is often reduced to clustering/classification of the genes/conditions. In contrast, candidate gene prioritization represents a unique process that can not be achieved with regular classification or clustering processes.
3. Many algorithms exist for classification and clustering, and most of these tools are actually implementing traditional techniques. The process of prioritizing, *i.e.*, ranking, genes with respect to a biological process of interest is rather new. It is therefore interesting to investigate whether advanced machine learning methods that have been developed only recently in academia can efficiently and accurately perform gene prioritization.

A second category contains the biological knowledge bases such as Ingenuity Pathways Analysis® and MetaCore™ GeneGo. These databases are very useful since they contain high quality genomic data which is in most of the cases manually curated by experts in the field. For instance, Ingenuity Pathways Analysis® eases the browsing of the scientific literature by providing manual annotations of the papers. MetaCore™ GeneGo proposes a module to visualize the results of your own experiments in a pathway context. Their main drawback is however that they represent passive knowledge bases. Gene prioritization can add significant

value to this field since the knowledge bases can be used to infer new associations (predictions), or to benchmark the approaches.

A third strategy related to gene prioritization is Gene Set Enrichment Analysis (GSEA), in which a set of genes is also investigated through the use of multiple data sources. However the goal of the GSEA strategy is to investigate and to characterize a complete gene set, without analyzing the individual genes in isolation. For one gene set, a GSEA will return a set of features, coming from multiple data sources, that correspond to molecular pathways and gene functions that best characterize the entire gene set. In addition, several GSEA tools are performing clustering or classification within the gene set [62, 123]. The main difference with gene prioritization is that, gene prioritization identifies which genes are the most promising candidates while a GSEA identifies the global function of the gene set and the corresponding pathways. These two strategies are complementary and, in fact, the first step of our gene prioritization strategy is the modeling part and is very similar to GSEA.

Proposed strategies

The present thesis describes the development of two distinct algorithms for gene prioritization that can both be classified as novelty detection methods. The first one is using basic statistics and is described in chapters 3, 4, 7, and 8. The second one is using a more advanced machine learning strategy and is described in chapters 5 and 6.

The first algorithm is based on simple statistics, accepts two inputs, and produces one outcome. The two inputs are, on the one hand, the genes known to be associated to the process of interest (the training genes), and on the other hand, the candidate genes to prioritize. The aim is to rank the candidate genes from the most promising genes on top to the less promising genes at the bottom, a three steps algorithm has been defined to do so. In the first step, the model is trained. More precisely, simple statistics are applied to the genomic data of the training genes, for instance, for annotation based data sources, a GSEA is performed in order to detect the most relevant ontological terms, the ones that best characterize the gene set. For most of the vector based data, the profiles of the training genes are collected and averaged, the averaged vector then represents the model of the training set. In the second step, the candidate genes are scored and ranked accordingly using the models built in the first step. For vector based data, the cosine of the angle between the averaged profile and the candidate gene profile is used as a score for that candidate. This second step results in a set of rankings, one per data source, that contain the most promising candidate gene at the top and the less promising ones towards the bottom. In the final step, the rankings are fused using the Order Statistics (OS), which corresponds to a late integration scheme. This results in a global

ranking with, again, the most promising genes at the top. This strategy is using basic statistics to build the models and therefore better models could theoretically be obtained with more advanced machine learning techniques. This method is described on figure 1.4 and in appendix A, it is further discussed in chapters 3 and 4.

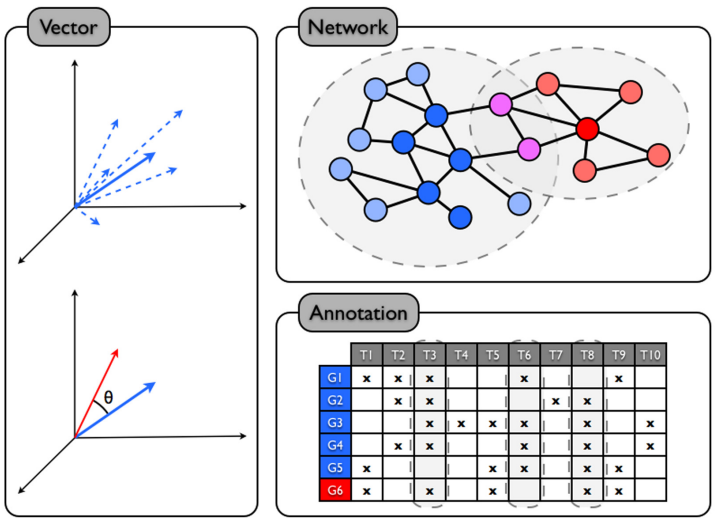

Figure 1.4: A basic model for gene prioritization that is based on simple statistics. For the three main data types, the training and the scoring schemes are described, the training information is plotted in blue and the candidate gene information is in red. (Vector - left panel) The training is performed by calculating an average vector (bold blue) from the training vectors (dotted blue). A candidate gene is scored by calculating the cosine similarity, denoted θ, between its vector (red) and the average vector calculated in the first step. A low cosine value indicates that the candidate gene profile is similar to the average profile. (Network - top right) In the training step, the subnetwork that contains the training genes and their direct partners is gathered (blue and purple nodes - large grey ellipse). The score of a candidate is based on the percentage of overlapping nodes between its own network (red and purple nodes - small grey ellipse) and the training network (the two purple nodes in this example). The larger the number of overlapping nodes compared to the total number of genes, the better. (Annotation - bottom right) For training, the annotation terms that are over-represented in the training set compared to the genome are kept for the second step (they are indicated by grey dotted rounded boxes in this example). Each term is associated to a p-value that represents the quality of the over-representation. A candidate is scored by combining the p-values of the annotated terms that have been kept in the first step using Fisher's omnibus. A more detailed description is given in appendix A.

The second approach presented makes uses of kernel methods. It means that only the algorithmic part is different, the inputs and outcomes are the same. First all the data sources are transformed into kernels (*i.e.*, matrices that contain the distances between the genes pairwise). Then, a one-class SVM algorithm is trained using simultaneously multiple kernels that correspond to multiple data sources. The training involves the maximization of a margin M so that on the hyperspace defined by the data, the training genes are separated from the origin (M represents the distance between the origin and the hyperplan that separates the training genes from it). Our implementation uses a soft margin to allow for a few misclassified data points. The SVM model is then used to score the candidate genes and rank them accordingly. Once again, the most promising genes are ranked at the top. The advantage of this technique is that each source is first transformed into a kernel which makes possible the merging of expression data and text mining data with minimal effort. The main difference with the previous method is that the integration happens during the modeling step (intermediate integration). This method is described in figure 1.5 and discussed in chapters 5 and 6.

1.3 Genomic data sources

The data sources are at the core of every bioinformatics approach, they are the basis upon which the algorithms derive novel hypothesis that when experimentally verified reinforce our knowledge. Gathering and analyzing the data therefore represent critical first steps of any bioinformatics method development. The amount of genomic data available has started to grow exponentially since the human genome was first drafted [131]. There are nowadays a plethora of databases that collect different types of data for different purposes. This section proposes a brief overview of what is available regarding our gene prioritization strategy.

The candidate gene prioritization problem focuses on genes, the data sources to consider are then also gene centric or gene product centric (mRNA, proteins). This means that other types of genomic data such as for instance patient centric data that are often used in disease marker discoveries [213, 21, 19] or in disease subtype classification [58, 84] have not been considered. The inclusion of this type of data is further discussed in chapter 9.

Several gene features can be retrieved including their functions, their expression profiles, their regulatory mechanisms (*e.g.*, transcription factors, miRNA), their sequences (*e.g.*, raw DNA/RNA/protein sequences, 2D/3D structures), their roles in biomolecular pathways, their associations with chemical components (including drugs), and their 'literature' (*i.e.*, what is written about them in the scientific literature). There exist several databases for each of these features, meaning that in total, it is a large amount of data to retrieve, analyze, organize and integrate. Typical data integration problems such as unbalance in data sources size, overlap

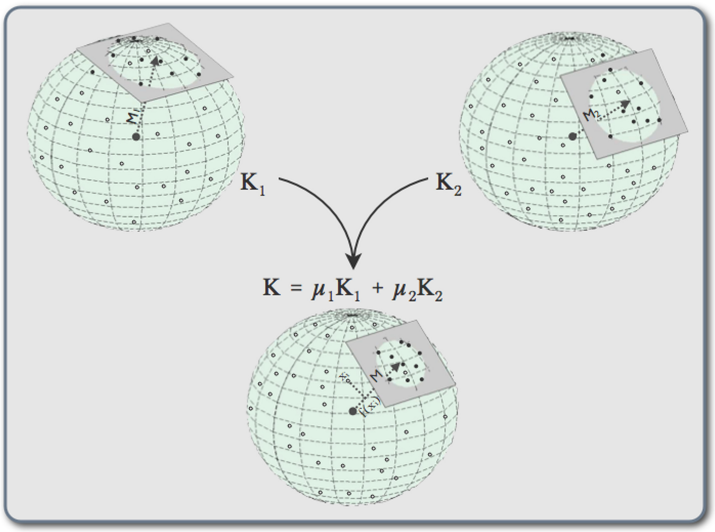

Figure 1.5: A more advanced model for gene prioritization based on a one-class SVM strategy. (Top-left) Schematic representation of the hyperplane (in grey) separating the (positive) training genes (filled circles) from the origin, along with the unlabeled genes (open circles). The larger the margin M, the better. (Top-right) Similar representation for a second kernel, with a different margin. (Bottom) The optimal convex combination of two kernels leads to a new kernel, where the margin between the positive genes and the origin is larger. A candidate is then scored by projecting its profile x_i along the vector that characterizes the hyperplane, the higher the score $f(x_i)$, the better.

between data sources, noise, bias towards well studied genes are discussed in the following sections.

1.3.1 Data versus knowledge

In bioinformatics, the term 'data' often refers to passive and unorganized information and is opposed to knowledge that is structured information that can be applied [54]. Gene prioritization is a predictive method, and as such, relies on the use of both existing knowledge and raw data in order to make predictions that are both accurate (by relying on knowledge) and novel (by relying on raw data).

On the one hand, knowledge bases are collection of curated data that represent the state-of-the-art in one specific domain. The data is often manually curated, meaning

that experts in the field went through the data and a consensus representation was created. This process is of course expensive and time consuming and it is often more efficient to also rely partially on computational tools to help the curation (*e.g.*, DIP [263, 262, 264, 206]). The goal of the curation is to reduce the number of false positive points (*i.e.*, noise) in order to obtain high quality data. The curation process is always a tradeoff between the quality of the data (less false positive points included) and the amount of data kept in (more data points included). In addition, knowledge bases such as Kegg [118, 120, 119], MetaCore and Ingenuity are highly valuable for researchers since they represent gold standards that can be used to benchmark computational approaches.

On the other hand, data repositories contain large amount of raw data, meaning that the data was not curated nor analyzed and that the biological signal is possibly hidden among background signal / noise. Repository such as the Gene Expression Omnibus (GEO) [70, 27] and ArrayExpress [181] are huge collection of microarray expression datasets that need to be pre-processed and analyzed. Also, yeast two-hybrid assays (Y2H) have been used to produce large collection of predicted PPIs that may contain a significant number of false positive [105, 104].

1.3.2 Primary and secondary data

A distinction is often made between primary data and secondary data [83]. On the one hand, primary data represents data relevant to the problem currently under investigation and is therefore case specific. For gene prioritization, it is the training data, in our case a set of known disease genes for the disease under study. A set of keywords or a dedicated expression dataset can be used alternatively for other prioritization methods [240, 248, 51, 169, 268]. On the other hand, secondary data is gathered beforehand and represents the field of investigation, and is therefore not case specific. For gene prioritization, the field is genetics and secondary data is the set of genomic data sources collected from various biological databases that describe the function of the genes and their roles in biological processes. The next sections describe in further details some characteristics of the secondary data sources.

1.3.3 Unbalanced data sources

The data sources differ not only by their content but also by their intrinsic properties. One property is the amount of data available per gene, if that amount varies between genes, then the data source is unbalanced and might be biased. This unbalance often reflects our current knowledge and is often observed between known genes that have been well studied over the years and almost unknown genes for which only few studies exist. An example is scientific literature for which well studied genes

are mentioned in many more publications than poorly characterized genes. At the contrary, there also exist data sources for which a stable amount of data is available per gene, they are unbiased. An example is a gene expression data set. Genome wide expression arrays measure the expression level of the whole transcriptome at once and therefore produce an unbiased output. This bias towards the known genes is observed in several of our data sources with sometimes a rather small effect as can be observed in figure 1.6. As expected, it is stronger for knowledge bases than for raw data repositories. The use of multiple sources for gene prioritization is therefore not enough to guarantee reliable novel predictions. On the one hand, the unbalanced data sources represent the current knowledge and should be used to obtain reliable results. On the other hand, the balanced data sources contain hidden knowledge and should be used to make novel predictions. For gene prioritization, the optimal strategy is to systematically use both types of data sources to leverage the effect between reliability and novelty.

1.3.4 Missing values

Another related property is the genome coverage, and consequently the missing value problem. This is a typical characteristic observed in many biological data sources. There are two scenarios, either the gene profile is missing completely or only some data points are missing. Although these two scenarios have different causes and consequences, similar strategies can be used among which the estimation of the missing values or the use a tailored calculation measure to take missing data points into account. It is often easier to estimate the missing values using dedicated algorithms (*e.g.*, replacing missing points by zero, k-nearest neighbors, local least square imputation or bayesian principal component analysis). For the knowledge bases, the missing value problem is directly related to the bias towards the known genes described above. The amount of data available per gene can vary, the extreme case is off course that nothing at all is known about a gene, then that gene is considered missing (see figure 1.6). For the raw data repositories, the amount of data available for each gene is stable. And although the technologies used are usually genome wide, there are always data points missing due to the technical limitations (*e.g.*, no probe spotted on the expression array for a gene). The data sources considered in the present work are also incomplete, we have circumvented the problem by developing a ranking method that take bias in to account or by estimating the missing values beforehand.

1.3.5 Multiple data sources

In engineering drawing, a three dimensional object can be represented by multiple two dimensional drawings, each one representing a view of the considered object (*e.g.*, front, left, right, top, bottom and rear views). A single view is usually not

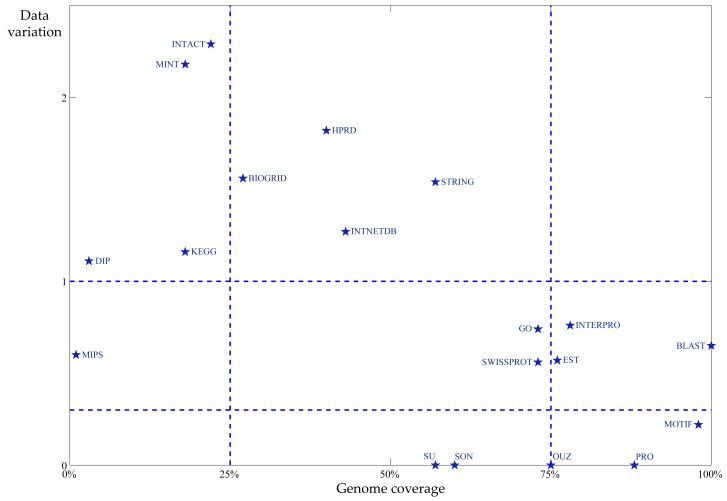

Figure 1.6: Bias in the data sources. The data sources are plotted on a two-dimension space, with the genome coverage on the x-axis, and the variation in the amount of data on the y-axis. Genome coverage is defined as the percentage of protein-coding genes for which data is available. Variation in the amount of data is defined as the standard deviation of data points (*e.g.*, number of annotated terms, number of interacting genes, number of samples) normalized by the mean of data points over the complete set. The blue dotted lines are plotted as guides to the eyes to discriminate the data sources. The data sources with lowest coverage are MIPS and DIP with less than 10% of the protein coding genes being present. This means that these data sources can only contribute to very specific problems. At the other end of the spectrum, Blast and Motif have the largest coverage since they are based on the gene sequences that are available for almost all protein coding genes. For similar reasons, the less biased data sources are sequences based (*e.g.*, Motif and ProspectR). Functional annotation data sources are moderately biased (*e.g.*, Gene ontology and SwissProt) when compared to the interaction sources such as Intact, Mint, and HPRD that are strongly unbalanced.

sufficiently informative to accurately describe the object while the use of multiple views in conjunction allows a much clearer definition of the considered object. One genomic data source can be seen as a single 'view' on the genome, and as for engineering drawing, one data source alone does not contain enough information to solve most biological questions. This is mainly because the molecular biology of the cell is not completely understood despite the massive amount of data available. Therefore, the integration of multiple heterogeneous data sources that can complete each others is believed to be more efficient than relying on a single source. This concept is crucial for any 'systems biology' approach, and also for the work described

in this thesis.

The overlap between different data sources has been studied by several research groups for human data [77, 153, 197, 134] and for model organisms (*e.g.*, worm, fly or mouse) [79]. These analysis have shown that, usually, a very poor overlap is observed between several data sources. For example, Franke *et al.* reported that only 20 interactions were shared by Reactome [116], Kegg [118, 120, 119], BIND [22, 23], and HPRD [188, 160, 193] for a total of 55606 interactions [77]. Although a little higher, the overlap is also poor for any pair of data sources (maximum is between Kegg and HPRD with 1074 common interactions for a total of 39988 interactions). Franke et *al.* explain that this is largely due to the use of different technologies (*e.g.*, yeast-two-hybrid assays, affinity purification by mass spectrometry, synthetic lethal screens) that create different network types (*e.g.*, physical interactions, metabolic pathways, genetic interactions, regulatory networks) with different global objectives [77]. A similar study has been performed on our data and is shown in figure 1.7. The results show a global agreement with the results previously reported, the overlap between four protein-protein interaction data sources is small (875 interactions for a total of 47431 - 1.84%). The biggest pairwise overlap is observed for BioGrid and HPRD (43.28%), which means that BioGrid is almost fully included in HPRD. The second biggest pairwise overlap is observed between IntAct and Mint (21.69%).

Another advantage of the use of multiple data sources is noise reduction. The fusion of multiple noisy data sources will indeed increase the signal-to-noise ratio since the noise appears random when compared to the real biological signal [237]. This is further discussed in chapter 4.

1.3.6 Multiple species

The use of model organisms that are close enough to human in the phylogenetic tree, and that can be easily and quickly bred is an efficient way of extending our knowledge in genetics in general, and in human genetics in particular through careful knowledge transfer. The ability of geneticists to study model organisms such as fruit fly (Drosophila melanogaster) and mouse (Mus musculus) has greatly contributed to extend our knowledge of several human disorders (*e.g.*, alzheimer's disease [276, 74] and diabetes mellitus [112, 40, 49]), to study key developmental pathways [89, 180, 177], and to produce unique biological datasets such as *in situ* fluorescence hybridization experiments in developing embryos [245].

The use of data sources from multiples species is not novel in bioinformatics. It was for instance used for sequence alignment problems to detect conserved regions between any two species [267] or to detect regulatory motifs / transcription factor binding sites in these conserved regions [5, 78]. It has also been widely used for gene expression data to analyze the co-expression patterns of the genes [145] in order to

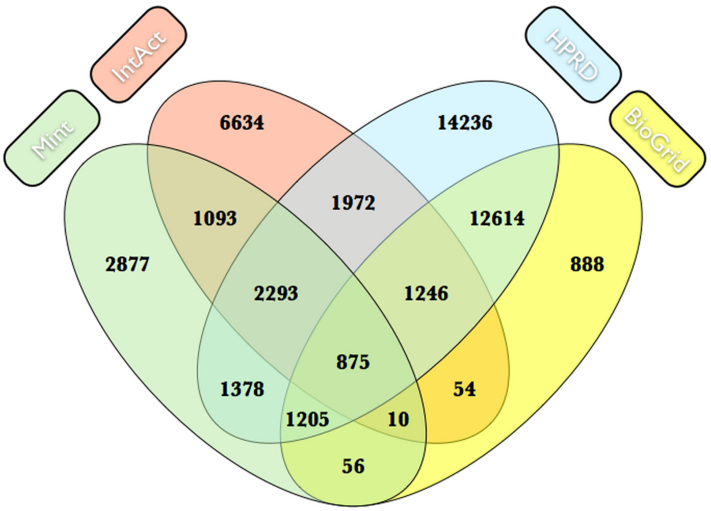

Figure 1.7: Overlap in interaction data sources. The overlap between four protein-protein interaction data sources is plotted. The numbers on the figure indicates the number of interactions shared by several data sources. Interactions are summarized at the gene level (different proteins from a single gene) and direction of the interactions is not taken into account ('A interacts with B' is similar to 'B interacts with A'). There are a total of 47431 interactions, among which 875 are shared by the four sets.

strengthen the further analyses such as motif detection [261, 161, 99] or human diseases studies [251, 149]. The use of multiple species data sources is however still an emerging topic for gene prioritization. Only few studies have been performed, and mainly focused on expression data [50, 106, 175, 179, 214, 14, 152]. Furthermore, the proposed approach was often disease specific through the integration for instance of mouse and human expression data sets that examine the insulin resistance associated with T2D pathogenesis [212]. The model proposed and discussed in the present dissertation is the integration of multiple data sources (not restricted to expression data) that cover multiple species in order to enhance candidate gene prioritization for human (see chapter 6).

1.3.7 Data type and similarity measures

This section describes different data types and the associated algorithms. Examples and details can also be found in figure 1.4.

Vector based data

With vector based data, each gene is characterized by a vector of data points. An example is gene expression data for which the expression level of genes are measured in several conditions/tissues leading to the generation of one expression vector per gene. Computing the similarity between two vectors can be achieved though the cosine similarity or the Pearson correlation coefficient among other techniques. Creating a model from a set of vectors can be achieved through averaging the vectors or by selecting a set of representative vectors (*e.g.*, via Principal Component Analysis - PCA). More advanced techniques such as bi-clustering or clustering can also be used to make the best of the data, and has proved to be more efficient for complex and noisy data such as expression data [64, 266, 140].

Network based data

Gene or protein networks are networks in which the nodes are genes or proteins and in which the edges represent an interaction between two nodes. A network is a representation that is very often used in bioinformatics due to its easy interpretation. However, different technologies and different aims will lead to networks that share only limited informations [77]. There are numerous network based methods available in order to make the best out of the data. For instance, the shortest path algorithm can be used to determine how close two nodes are within a network, alternatively, the overlap between the two neighborhoods can also be compared. For a given gene, measures such as centrality or connectivity can also assess whether that gene is key for the network under consideration [85, 20, 143, 8].

Annotation based data

Ontologies are vocabularies whose terms describe, for instance, gene functions (*e.g.*, Gene Ontology - GO [17]), disease phenotypes (*e.g.*, Mammalian Phenotype Ontology - MPO [218, 217]), or species anatomy (*e.g.*, Foundational Model of Anatomy - FMA [204]). Most ontologies are structured trees so that root terms are general terms and that leaf terms are more specific. Most ontologies also allowed multiple relationship between parent nodes and child nodes to reflect the complexity of the modeled system. Most of the annotation based data sources are binary (*(e.g.*, Gene Ontology), but some are also associated to a score that assesses the reliability of the association (*e.g.*, Sequence Ontology - SO [71, 72, 165]). Several algorithms have been developed to work with ontologies in order to predict novel associations (*e.g.*, FuncBase [29], PoGo [117]) or to find the more representative terms for a set of genes (*e.g.*, DAVID [62, 103]). As already described in section 1.2.2, modeling can be achieved through a GSE analysis to detect the most representative features.

Sequence based data

Each gene has a DNA sequence, potentially several RNA and protein sequences. Several tools have been developed to compare sequences (mainly through alignments), to identify functional and regulatory elements that reside within them, and to predict their 2D/3D structures. Tools such as BLAST (Basic Local Alignment Search Tool) and FASTA [184] are based on sequence alignment algorithms that allow the comparison of DNA, RNA and protein sequences in order to detect homologous sequences. They can also be used to estimate the similarities between all proteins pairwise as done in the SIMAP project [16, 198, 199].

Identifying the regulatory elements residing within DNA sequences is a key challenge in bioinformatics since understanding when, how and where a gene is active helps in building the cellular functional map. To this end, several computational approaches have been developed. For instance, putative Transcription Factor Binding Sites (TFBS) can be identified using Toucan that uses the upstream sequences of the protein coding genes together with human-mouse conservation [5, 6] among other tools [194, 61, 141, 98]. Other key players in regulation are miRNAS that have been recently the focus of bioinformatics tools that try to identify the putative miRNA binding sites [10, 138].

Similarly identifying the functional elements within sequences has received a lot of interest from the bioinformatics community, and tools such as InterProScan [274] are now able to predict accurately the protein functional domains from the protein sequence. All these tools can be used to mine sequences and to derive more structured data that are either annotation or vector based data.

Kernel data

An alternative method is the use of kernel methods that approach the problem by mapping the data from the original feature space into a high dimensional space, in which each coordinate corresponds to one original feature. This mapping is in fact a trick, termed the 'kernel trick', that allows the use of a linear classifier (*e.g.*, SVM) in the mapped space, which would not have been successful in the original space. The linear classification in the mapped space is equivalent to a non-linear classification in the original feature space. The best advantage of kernel methods is that it is in fact not necessary to find the mapping function and therefore to compute the exact coordinates of the data in the mapped space. The computation of the similarities between the objects (*e.g.*, through the inner product between their profiles) creates a kernel matrix and makes sure that the mapping function does exist (without actually determining it).

In this case, the gene prioritization problem can be defined as a novelty detection problem, formally a one-class SVM, for which only the positive training genes

are modeled in order to rank the unlabeled genes. For training, the approach we propose finds a hyperplane separating the positive data from the origin. For scoring, the distance between the hyperplane and the projection of the candidate gene profile along the direction of the hyperplane is used. Figure 1.5 summarizes schematically the kernel approach.

1.4 Validation

1.4.1 Benchmarking

A key step in algorithmic development is the benchmarking of the approach in order, first, to validate the global strategy and to get an estimate of the performance on real data, and, second, to compare it with existing methods. An easy and common way of benchmarking a predictive algorithm is the cross-validation procedure. In a cross-validation setup, a proportion of the existing knowledge is used for training while the remaining part is used for testing. This is usually repeated a number of times with different repartition of the existing knowledge between training information and testing information, so that the results do not depend on a single repartition. The predictions are then compared to the expected values and the accuracy of the algorithm is estimated by measuring how correct the predictions are. The proportion of existing knowledge allocated for training can vary, for instance, in a 10-fold cross-validation, 90% of the data is used for training and 10% for testing. The most extreme case is the leave-one-out cross-validation (LOOCV) in which a single data point is reserved for testing. For gene prioritization, the existing knowledge is represented by the genes known to be involved in a collection of genetic diseases. LOOCV is usually preferred since it simulates the situation in which a single novel disease gene is discovered using all currently known disease genes. For one LOOCV iteration, one of the known disease gene is left out while the remaining genes are used for training. The left-out gene is then mixed up with randomly selected genes to build the candidate set of a given size. Alternatively, the nearest chromosomal neighbors can be used instead to mimic a positive locus. The candidate set is then scored and a ranking of the candidate genes is recorded. If the algorithm is working perfectly, the left-out gene ranks first since its implication in the process under study is in fact already known. This is repeated until every known gene has been, in turn, left-out. A schematic LOOCV procedure is presented in figure 1.8. For classification problems, a cross-validation procedure is usually followed by the a Receiver Operating Characteristic (ROC) analysis. For gene prioritization, the results of a single run is not a classification but a ranking of the candidates. However, it is possible to derive a binary classification by applying a threshold on this ranking, and thus to create a point in the ROC space. The ROC space is defined by the False Positive Rate (FPR, also one minus specificity) on the x-axis, and the True Positive Rate (TPR, or sensitivity, recall) on the y-axis.

By varying this threshold, a complete ROC curve can be created. The goal in ROC space is to be in the upper-left-hand corner (maximum TPR for a minimum FPR). The Area Under the ROC Curve (AUC) is often used to summarize the ROC curve and estimate how well positives are retrieved on average. More details can be found in figure 1.8.

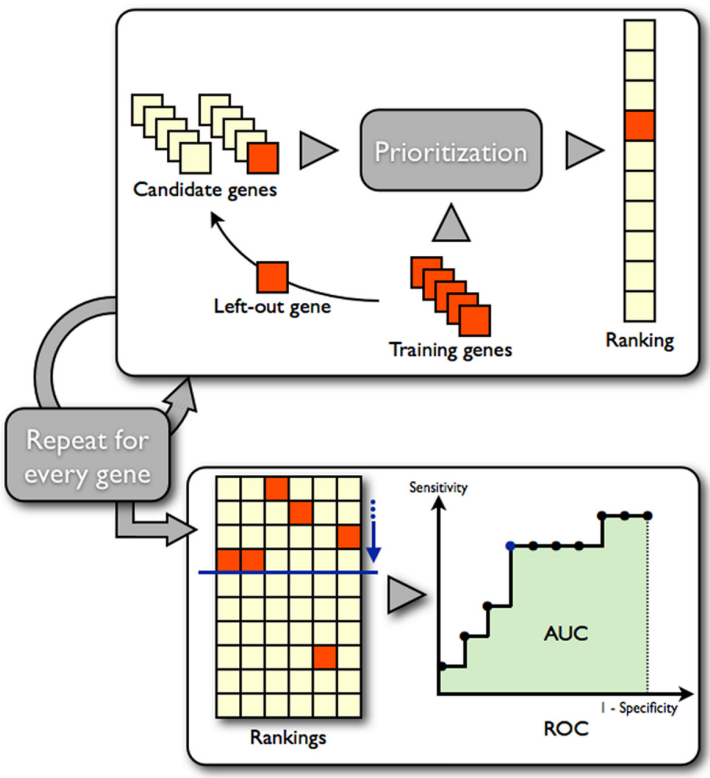

Figure 1.8: The leave-one-out cross-validation procedure. The validation consists of a repetition of prioritization runs. For each run, one of the training genes (orange boxes) is left-out and mixed with candidate genes (light yellow boxes). The remaining training genes are used for training and, after prioritization, the ranking of all the candidate genes (including the left-out gene) is recorded. This step is repeated so that all the training genes are, in turn, left-out. In a second step, these rankings are used to build a ROC curve. By using a threshold on the matrix (blue line), it is possible to define a binary classifier, the associated sensitivity and specificity, and therefore to draw a point in the ROC space. By varying the threshold along the matrix, it is possible to determine a complete ROC curve and its AUC.

A benchmark analysis relies on one, but preferably several gene set to validate. Each gene set must represent a precise biological process, a function, or a genetic disease. Therefore, most of them are either manually built or derived from knowledge bases that contain reliable disease, pathway and functional information. A leave-one-out cross-validation as defined above will then estimate how easy it is to retrieve one of the genes when using all the other to modelize the underlying process. Since the prioritization methods rely on the 'guilt-by-association' concept, cross-validation also assesses whether the genes within the set are similar to each other, therefore meaning that the entire set is homogeneous. The validations described in chapters 3, 4, and 7 are indeed using OMIM, Gene Ontology, Ingenuity and Metacore among other sources to build the validation sets.

Many pitfalls are inherent to the cross-validation approach, the main disadvantage is the optimistic performance estimation [167]. The LOOCV procedure is measuring the ability of an algorithm to capture what is already known by pretending it is not known. For some problem, it is however not sufficient since the underlying data might contain explicitly the same information. For gene prioritization, some of the underlying data sources that are at the core of the prioritization do contain information about the gene-disease association, which eases the retrieval of the left-out genes. To circumvent that problem, one possibility is to discard the data sources that might contain explicitly the gene-disease associations in order to get a better estimate of the real performance. It is however not an optimal solution since a potentially interesting data source might be discarded only because it contains explicit gene-disease associations among other information. Another solution is to use rolled back data, that is data prior to the discovery of the gene-disease association in order to benefit from the data source without including directly the gene-disease association. However, rolling back genomic data is costly and although it has already been used for gene prioritization (for literature data only [4]), it is not yet common in bioinformatics.

A main disadvantage of the AUC is that it aggregates the performance across the entire curve. It is sometimes more interesting to look deeper at the beginning of the ROC curve in order to estimate how well the algorithm is working for the top predictions. To circumvent that, the notion of partial AUC has been defined as the AUC below the ROC curve after truncation [250, 65]. This allows the measure of the performance on the top predictions only but have rarely been used to estimate the performance of the gene prioritization methods, the full AUC is still preferred. An alternative to ROC curves are Precision Recall curves (PR) that are defined by the True Positive Rate (TPR, or sensitivity, or recall) on the x-axis, and the Positive Predicted Value (PPV or Precision) on the y-axis. In PR space, the goal is to be in the upper-right-hand corner (maximum TPR and maximum PPV). The main difference is that PR curves are not using the true negatives, that is the negative predictions that are indeed negative. A direct consequence is that PR curves give a more informative picture of an algorithm's performance when dealing

with highly skewed data sets [36, 41, 86, 127, 215]. Also, looking at PR curves can expose differences between algorithms that are not apparent in ROC space [59]. PR curves appear thus more suitable for gene prioritization strategies since there are more negative genes (randomly selected genes) than positive genes (known disease genes), and since negative genes are not the focus. However, most of the benchmark results of the existing prioritization strategies are still reported as ROC curves and not PR curves. There exists other alternatives such as cost functions [67, 68] that can take into account the cost associated to a false positive and a false negative. Obtaining a reliable estimate of the real performance is an interesting problem but it is often not the main concern since this is only an estimate of the real performance. When possible, it is often more interesting to apply the strategy to real biological questions to experimentally validate it.

1.4.2 Experimental validation

Benchmarking can prove that the overall method is correct and can be used to estimate the performance of the approach. However it only represents the first step of the validation, the second being the experimental validation, that is the application to real biological problems. Starting from a biological process of interest, and possibly from candidate genes, predictions are made *in silico* using prioritization strategies, and only these predictions are then experimentally validated. Since experimental validations are costly and time consuming. Several studies that make predictions do not validate them experimentally but rather estimate how promising are some of the candidate by examining the existing literature and the publicly available results of independent experiments [234, 229, 73].

The possible experimental validations very much resemble the three situations described section 1.2.1 (chromosomal aberration in a patient with a genetic condition, differential expression of genes in a disease tissue, linkage analysis). The inclusion of gene prioritization in wet lab based workflows is presented in chapter 3 and 8.

1.4.3 External validation

Two main goals of bioinformatics are (i) the definition of computational methods that use existing knowledge (including data) to create additional knowledge (hypothesis) and (ii) their integration in research workflows to solve real biological problems. To allow such integration, it is crucial to go further than the development of conceptual methods and to implement these approaches into publicly available tools that can be used by bioinformaticians and biologists world wide. Several technologies and IT models can be used to make the approaches available to the public, an overview of the models chosen for this thesis is described in figure 1.9.

The proposed architecture allows multiple interfaces using multiple programing languages, the use of the tools from the command line or using GUIs facilitating thus its inclusion within workflows.

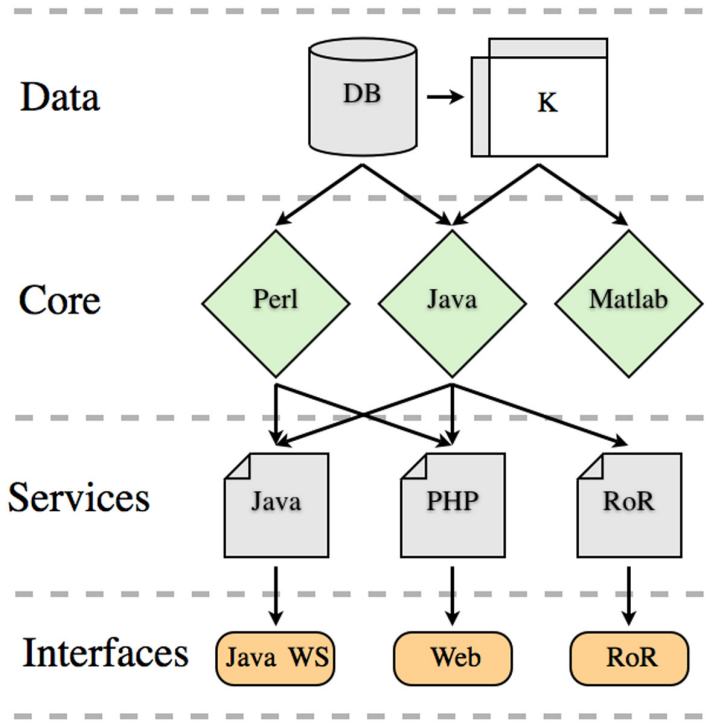

Figure 1.9: The IT system behind our tools. At the data level, two different objects: the raw data in the database (for the basic statistics version) and the kernels (K) that are created from the raw data (for the kernel based version). The core level contains all the necessary code to perform gene prioritization. The version based on basic statistics is written in Perl and in Java. The kernel based version is written in Matlab and Java. All the core modules can be run on the command line to perform candidate gene prioritization very efficiently. At the services level, web services are available publicly in Java, PHP and Ruby on Rails (RoR). When available, the services can interact with both Perl and Java cores depending on the task. These can be accessed by computer programs to include prioritization within computational workflows. At the client level, several interfaces have been developed including a Java Web Start client (for the basics statistics version only), classic websites (for both versions), and a Ruby On Rails client (for the basics statistics version only with a special emphasis on miRNA prioritization). These three clients are the ones that are mostly used by biologists and geneticists (see chapter 4 for an overview of the user base).

An important way to bring these computational methods to researchers is to create web tools that are easy to use. This statement implies that the tool interface should be intuitive so that potential users easily find their way and achieve what they want to achieve. Another feature is the development and the maintenance of tools and documents to help and guide users (this can include for instance a mailing list, a FAQ section, and a manual together with running examples). The chapters 4 and 6 present the efforts made towards the development of such interfaces.

Bioinformaticians are very much interested in integrating different tools together and in querying these tools automatically, which a web interface does not always allow. In the recent years, web services have emerged as an efficient option to grant access to computational tools. Also, tools such as Taverna [172, 107], Kepler [11], and UCSC Galaxy [228, 35] allow the creation of elaborated computational workflows that combine several web services together. Chapter 4 also presents the development of such web services.

1.5 Thesis overview

Chapter 2 introduces the scientific topic of the present dissertation. More precisely, it is a review of the numerous gene prioritization methods that have been developed in the last decade. It focuses on the tools that are web based, freely accessible, and available for human, but other approaches are also discussed.

Chapter 3 introduces a first prioritization strategy based on order statistics. It also describes the implementation of this method into a publicly available Java based tool termed 'Endeavour'. In addition, it also presents the validation of the approach through benchmarking and cross-validation on known gene sets. More important, the integration of Endeavour into an experimental workflow is described and applied to the analysis of an atypical DiGeorge syndrome deletion.

Chapter 4 presents a major update of our tool Endeavour. The first improvement is the support of multiple species through the addition of three model organisms (mouse, rat, and worm) and their corresponding data sources. The second improvement is the development of a web based interface, easier to use, to supplement the java based interface that requires more expertise.

Chapter 5 is a description of a second prioritization strategy. This strategy is based on kernel methods and is using a 1-SVM algorithm. It outperforms our previous method based on order statistics using the same benchmark data.

Chapter 6 describes the implementation of the method described in chapter 5 into a web based tool termed 'MerKator'. In addition, it describes a major improvement: the cross-species module, that is the possibility to prioritize genes from one species using data from multiple other species.

Chapter 7 presents several large-scale benchmark analysis performed to validate the approach behind Endeavour. The benchmarks are based on knowledge bases such as Ingenuity Pathway Analysis, MetaCore GeneGo, Gene Ontology, and Kegg.

Chapter 8 describes the development of a gene prioritization tool dedicated to fruit fly termed 'Endeavour-HighFly', and its experimental validation through the study of the atonal mediated neural development. In addition, this chapter contains a collection of experimental validations of Endeavour.

Chapter 9 presents some conclusions about the work described in the present dissertation. Based on these conclusions, it also elaborates on several research avenues to be explored in the future.

Chapter 2

A guide to web tools to prioritize candidate genes

2.1 Summary

In the last decade, the gene prioritization problem has received a lot of attention from the bioinformatics community. Several approaches have been defined, benchmarked, and implemented into computational tools [275, 2, 110, 4, 50, 273, 211, 268, 205, 152, 51, 240, 102, 248, 272, 82, 265, 126, 196, 80, 163, 148, 39, 239, 77, 235, 169, 236, 187, 186]. These computational tools differ by the inputs they accept, the outputs they generate, the genomic data sources they integrate, and the prioritization strategy they use. Furthermore, several of them have been experimentally validated and have proved to be useful in finding novel disease or pathway genes. There was however only few gene prioritization reviews, and their authors mostly made use of only three to five tools to predict new disease contributing genes [234, 73]. Furthermore, there were no catalog of tools from which users could select the tools that suit best their needs. This review is a remedy to that problem and proposes a website that contains enough information for the reader to select the web based prioritization tools he wants to investigate more. The website represents a dynamic version of the static review and is meant to be updated on a regular basis.

Briefings in Bioinformatics Advance Access published March 21, 2010

doi:10.1093/bib/bbq007

A guide to web tools to prioritize candidate genes

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

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

Abstract

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

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

BACKGROUND

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

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

*These authors contributed equally to this work.

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

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

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

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

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

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

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

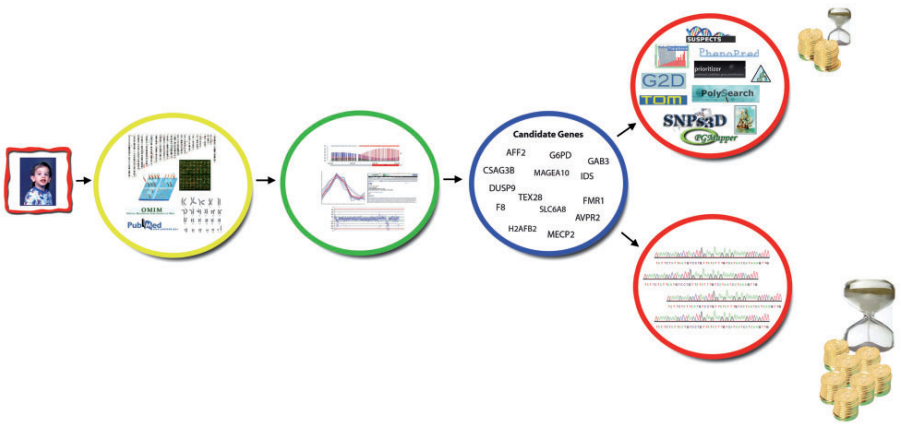

Figure 1: A major challenge in human genetics is to unravel the genetic variants and the molecular basis that underlay genetic disorders. In the past decades, geneticists have mainly used high-throughput technologies (such as linkage analysis and association studies). These technologies usually associate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

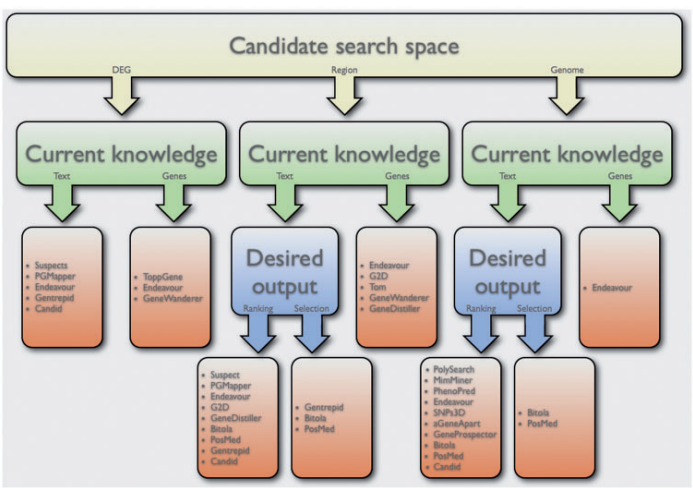

Figure 2: Decision tree that categorizes the 19 gene prioritization tools according to the outputs they use and the outputs they produce. This tree is designed to support the end users in their decision so that they can choose the tools that suit best their needs. By starting from the first question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

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,

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

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

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 1.

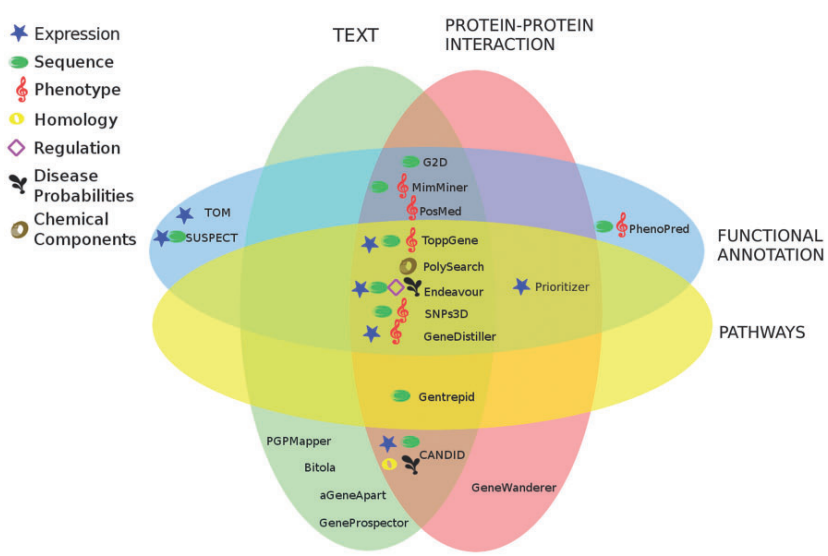

Figure 3: Repartition of the 19 tools according to the data sources they use. The four data sources most commonly used are Text (functional and interactions mining), protein–protein interactions, functional annotations and pathways and are therefore represented as large ellipses. The additional seven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

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

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.

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.

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.

2.2 Contribution of the PhD candidate

The PhD candidate has reviewed the literature in order to define the list of the gene prioritization methods to include in the publication. The PhD candidate has reviewed one third of the associated web tools, and mined the literature to find experimental validations. The PhD candidate has created the associated website and has written the manuscript.

2.3 Discussion

In this paper, the discussion of the tools' characteristics has allowed us to divide them up into several categories. One can, for instance, clusters the tools according to the inputs they need or to the data sources they use and therefore select the ones that fit best his/her needs. For this review, only the 18 prioritization tools that were publicly available as web servers were retained. This set of 18 tools is a subset of all the prioritization approaches that have been proposed since many of them have not been implemented into publicly available computational tools, or are provided as stand alone scripts / code to download. This selection decision was motivated by the fact that the end users of these tools are mainly biologists who are more willing to use web servers than scripts (not mentioning the implementation of a method from scratch). However, several of the discarded methods were based on interesting prioritization concepts or algorithms that are often not found in the existing web tools. It was therefore decided to include them in the web portal so that users could still consider the use of a method for which no web server is available. Upon this decision, the portal was then extended with 6 additional methods which brings the total to 24 prioritization strategies. The website was also extended with recent publications that describe the application of some of the existing tools to biological questions. Last update, a search engine was built to help users to find the tools that fulfill his/her criteria.

2.3.1 Assessing the relevance of the predictions

Creating a portfolio of these web tools and maintaining the corresponding website is only the first step towards a better understanding of this especially fast evolving field. A number of independent studies have been performed to identify the most promising candidates for type 2 diabetes mellitus and obesity [234, 73], however these predictive studies were not followed up by experimental studies to estimate the quality of the predictions, and they were only focusing on four or five prioritization tools at the time. Altogether, this means that a critical assessment of the tools' performance on real biological cases is still missing. One possibility is

to systematically estimate their ability to predict disease-gene associations as they are reported in the literature but before their inclusion in genomic databases (a strategy similar to the CASP strategy [164]). A second possibility is to start with the prediction of interesting candidates for many genetic disorders in a first time, and to continue, a few years later, with a follow-up study to check the results. This strategy expands over several years which makes it difficult to realize in practice, an alternative is the use of rolled back genomic data. With five years rolled back data, disease-gene associations from the last four years can be assessed directly without having to actually wait five years (to leave a one year gap). Rolling back data is however only an emerging research interest in bioinformatics, and many databases are not keeping data for that long, which does not ease the process. Furthermore, the data would ideally have to be rolled back to a different time for every association making the validation of hundreds of genes painful. The chapter 3 describes a small rolled back validation of Endeavour in which the rolled back literature data is used to prioritize disease genes. It shows that, it would have been possible to predict the association using literature data 6 months prior to discovery (three out of nine monogenic disease genes among the top 2% candidate genes).

Following the publication described in this chapter and the conclusions of our preliminary study (described in the chapter 3), it was decided that a critical assessment of the tools over a rather short period of time (several months) could be realized. The two other possibilities should however not be discarded since they also represent nice opportunities to estimate which prioritization methods are more efficient than others. The core concept is to use the disease-gene associations from major human genetics journals as soon as they are published and to check whether or not the existing tools could have predicted them. A main difference with the CASP project is that the experiments are not run by the people who have developed the tools but by our bioinformatics team. This involves that a fine tuning of the tools is excluded, including for the tools developed internally. Another difference is that the disease-gene association is not known beforehand and kept secret as the protein 3D structure is for the CASP experiments, meaning that the prioritization experiments have to be run as soon as the publication is made public (advanced publication for most journals). A critical assessment of several gene prioritization tools is currently realized internally in collaboration with Daniela Nitsch and Francisco Bonachela Capdevila.

For Endeavour, the preliminary results show that 39 among the 43 associations could have been predicted using a threshold of 30%, and 18 if a more conservative threshold of 10% is used. The median over all 43 associations is 11.21%. These results are a little bit lower than the benchmark results, and correlate with the observation of Myers *et al.* that the performance observed in cross-validation studies is likely to be higher than that observed in prospective studies [167]. Another possibility is related to the complexity of the disorder. Nowadays most of the novel disease genes that are reported are associated with complex disorders for

which accurate predictions are more difficult to make than for disorders with simple mendelian inheritance pattern. This is mainly due to the fact that our approach relies on the similarity between candidate genes and the known disease genes. For a mendelian disorder, this hypothesis has shown to work very well since the disease genes are often acting together in a protein complex (*e.g.*, Usher syndromes). For a complex disorder, it is however not a single complex, nor a single pathway that is disturbed, and the known genes are sometimes active in different processes that when perturbed are leading to the same type of disease. This is making the predictions for complex disorders less accurate although still a lot better than random (see chapter 3). Last point, it might also be the case that a fine tuning of Endeavour (*e.g.*, selection of the optimal data sources through cross-validation) is required to make the best predictions. Still, Endeavour remains as one of the best tools among the tools assayed in the study and this work will serve as a basis for a large predictive study. In addition, the results do not rely too heavily on the text data source since the performance when 'Text' is excluded remains similar (median rank ratio 11.21% using all data sources, 13,19% when text is excluded). This illustrates that literature is not the only informative data source and that accurate predictions can be made without considering it.

Chapter 3

Gene prioritization through genomic data fusion

3.1 Summary

This paper describes the development of a gene prioritization strategy, its implementation into a software termed Endeavour, and its validation. It is therefore representing the first step of the work described in the present dissertation. It was published in Nature Biotechnology in May 2006.

The proposed strategy is based on the assumption that the most promising candidate genes are the ones that exhibit similarities with the genes already known to be involved in the process under study [115, 219, 87]. It uses a data fusion algorithm and multiple genomic data sources. The inputs of the methods are a set of genes known to be involved in the biological process under study (training set), and a set of candidate genes to prioritize (candidate set). The proposed algorithm is made up of three steps. The first step is the training step in which the known genes are used to modelize the process under study. That is, for each data source, the creation of a model using basics statistics. For instance, for an annotation based data source (*e.g.*, Gene Ontology), the model contains all the annotation terms that are over-represented in the training set compared to the genome. In our case, the over-representation is calculated using the binomial distribution. Next comes the scoring step in which the models built in the first step are then used individually to score the candidate genes. The candidate genes are then ranked according to their scores, resulting in one ranking per model. Within that ranking, the most promising candidate genes are located at the top, and the less promising at the bottom. The data fusion happens at the final stage. More precisely the rankings

are merged using the Order Statistics (OS), a method that takes missing data into account therefore allowing a leverage between well known genes and unknown genes. The output of the algorithm is a single global ranking of the candidate genes. More details can be found in appendix A.

The approach was benchmarked on known diseases and pathway sets but was also experimentally validated. The first validation is the analysis of regulatory genes for myeloid differentiation. Endeavour was used to prioritize candidate genes predicted to be upregulated during myeloid differentiation by a cis-regulatory model. The PCR results showed that 2 out of the top 20 genes predicted by the cis-regulatory model were upregulated while after prioritization, 8 out of the top 20 genes were upregulated. The second validation describes the exploration of an atypical deletion observed in DiGeorge syndrome patients and located on 22q11. The 58 genes that are located in that region were then prioritized using Endeavour. The candidate genes were then individually assayed by knock down experiments of zebrafish embryos. The YPEL1 knock down embryos exhibited physical characteristics that are compatible with the phenotypes of the DiGeorge syndrome patients. This *in vivo* validation suggested that YPEL1 is involved in craniofacial development and represents a promising candidate gene for the DiGeorge syndrome.

Gene prioritization through genomic data fusion

Stein Aerts[1,4,5], Diether Lambrechts[2,5], Sunit Maity[2,5], Peter Van Loo[3–5], Bert Coessens[4,5], Frederik De Smet[2], Leon-Charles Tranchevent[4], Bart De Moor[4], Peter Marynen[3], Bassem Hassan[1], Peter Carmeliet[2] & Yves Moreau[4]

The identification of genes involved in health and disease remains a challenge. We describe a bioinformatics approach, together with a freely accessible, interactive and flexible software termed Endeavour, to prioritize candidate genes underlying biological processes or diseases, based on their similarity to known genes involved in these phenomena. Unlike previous approaches, ours generates distinct prioritizations for multiple heterogeneous data sources, which are then integrated, or fused, into a global ranking using order statistics. In addition, it offers the flexibility of including additional data sources. Validation of our approach revealed it was able to efficiently prioritize 627 genes in disease data sets and 76 genes in biological pathway sets, identify candidates of 16 mono- or polygenic diseases, and discover regulatory genes of myeloid differentiation. Furthermore, the approach identified a novel gene involved in craniofacial development from a 2-Mb chromosomal region, deleted in some patients with DiGeorge-like birth defects. The approach described here offers an alternative integrative method for gene discovery.

With the advent of 'omics, identifying key candidates among the thousands of genes in a genome that play a role in a disease phenotype or a complex biological process has paradoxically become one of the main hurdles in the field. Indeed, contrary to some early concerns in the community that a lack of sufficient global data would still be a limiting factor[1], it is precisely the opposite, a bounty of information that now poses a challenge to scientists. This has translated into a need for sophisticated tools to mine, integrate and prioritize massive amounts of information[2,3].

Several gene prioritization methods have been developed[4–10]. Most of them determine, either directly or indirectly, the similarity between candidate genes and genes known to play a role in defined biological processes or diseases. These methods offer several advantages but also pose a number of challenges. Indeed, even though multiple data sources are available, such as Gene Ontology (GO) annotations[4–6,10], protein domain databases[6,10], the published literature[5,7], gene expression data[5,7,10] and sequence information[8–10], most of the available programs access only one or two of these databases, which each have their limitations. For instance, functional data sources (GO and literature) are incompletely annotated and biased toward better-studied genes[8], whereas sequence databases have thus far been used only to produce general disease probabilities[8,9]. Some of the existing tools access more than two databases, but do not provide an overall ranking that integrates the separate searches[5,10]. Several tools rank disease genes but only one of them prioritizes genes involved in biological pathways[10], and none offers the combination of both. Thus far, only two prioritization tools[5,10] are publicly available. Thus, there is still a need for improvement of gene prioritization.

Here, we report the development and characterization of a new gene prioritization method, and offer its freely accessible, interactive and flexible software[1]. Compared to existing methods, ours provides additional opportunities for candidate gene prioritization: it accesses substantially more data sources and offers the flexibility to include new databases; it provides the user control over the selection of training genes and thereby takes advantage of the expertise of the user; it prioritizes both known and unknown genes, ranks genes involved in human diseases and biological processes, and it uses rigorous statistical methods to fuse all the individual rankings into an overall rank and probability.

RESULTS

Principles of prioritization used by Endeavour

Genes involved in the same disease or pathway often share annotations and other characteristics in multiple databases. Indeed, genes involved in the same disease share up to 80% of their annotations in the GO and InterPro databases[6], whereas genes involved in a similar biological pathway often share a high degree of sequence similarity with other pathway members[11]. It is therefore reasonable to assume that this similarity among genes is not restricted to their annotation or sequence alone, but is also true for their regulation and expression. We reasoned that a bioinformatics framework capable of comparing and integrating all available gene characteristics might be a powerful tool to rank unknown candidate 'test' genes according to their similarity with known 'training' genes, and based on this notion, we developed Endeavour. Prioritization of genes using this algorithm involves three steps (**Fig. 1**). To validate its performance, we used several complementary strategies discussed below.

[1]Laboratory of Neurogenetics, Department of Human Genetics, [2]The Center for Transgene Technology and Gene Therapy, [3]Human Genome Laboratory, Department of Human Genetics, Flanders Interuniversity Institute for Biotechnology (VIB), University of Leuven, Herestraat 49, bus 602, 3000 Leuven, Belgium. [4]Bioinformatics Group, Department of Electrical Engineering (ESAT-SCD), University of Leuven, Belgium. [5]These authors contributed equally to this work. Correspondence should be addressed to S.A. (stein.aerts@med.kuleuven.be).

Published online 5 May 2006; doi:10.1038/nbt1203

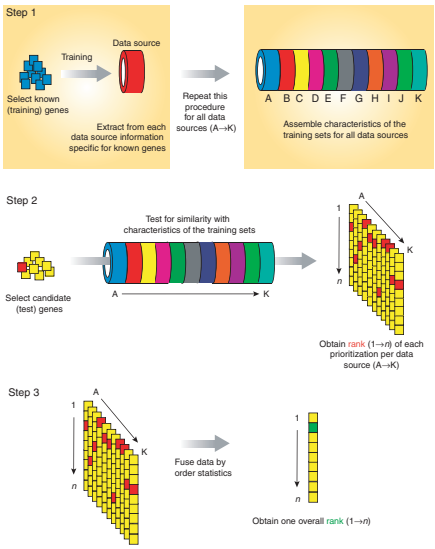

Figure 1 Concept of prioritization by Endeavour. Candidate test genes are ranked with Endeavour based on their similarity with a set of known training genes in a three-step analysis. In the first step (upper panel), information about a disease or pathway is gathered from a set of known training genes by consulting various data sources. Training genes can be loaded automatically (based on a Gene Ontology term, a KEGG pathway ID or an OMIM disease name) or manually. The latter allows the incorporation of expert knowledge. The following data sources are used: A, literature (abstracts in EntrezGene); B, functional annotation (Gene Ontology); C, microarray expression (Atlas gene expression); D, EST expression (EST data from Ensembl); E, protein domains (InterPro); F, protein-protein interactions (Biomolecular Interaction Network Database or BIND); G, pathway membership (Kyoto Encyclopedia of Genes and Genomes or KEGG); H, *cis*-regulatory modules (TOUCAN); I, transcriptional motifs (TRANSFAC); J, sequence similarity (BLAST); K, additional data sources, which can be added (e.g., disease probabilities). In the second step (middle panel), a set of test genes is loaded (again, either manually or automatically by querying for a chromosomal region or for markers). These test genes are then ranked based on their similarity with the training properties obtained in the first step, which results in one prioritized list for each data source. Vector-based data are scored by the Pearson correlation between a test profile and the training average, whereas attribute-based data are scored by a Fisher's omnibus analysis on statistically overrepresented training attributes. Finally, in the third step (lower panel), Endeavour fuses each of these rankings from the separate data sources into a single ranking and provides an overall prioritization for each test gene. As such, Endeavour prioritizes genes through genomic data fusion—a term, borrowed from engineering to reflect the merging of distinct heterogeneous data sources, even when they differ in their conceptual, contextual and typographical representations.

Validation of Endeavour when accessing individual data sources

For each individual data source, we assessed whether our approach is capable of prioritizing genes known to be involved in specific diseases or receptor signaling pathways. To this end, we performed a large-scale leave-one-out cross-validation. In each validation run, one gene, termed the 'defector' gene, was deleted from a set of training genes and added to 99 randomly selected test genes. The software then determined the ranking of this defector gene for every data source separately. We used 627 training genes, ordered in 29 training sets of particular diseases automatically selected from the Online Mendelian Inheritance In Man (OMIM) database (see **Supplementary Notes** online for selection procedure). For pathway genes, we compiled three sets of training genes involved in the WNT (43 genes), NOTCH (18 genes) and epidermal growth factor (15 genes) pathways. As a negative control for training genes, we assembled 10 sets of 20 randomly selected genes.

Thus, a total of 903 prioritizations (627 for the disease genes, 76 for the pathway genes and 200 for the random sets) were performed for each data source. From these, we calculated sensitivity and specificity values. Sensitivity refers to the frequency (% of all prioritizations) of defector genes that are ranked above a particular threshold position. Specificity refers to the percentage of genes ranked below this threshold. For instance, a sensitivity/specificity value of 70/90 would indicate that the correct disease gene was ranked among the best-scoring 10% of genes in 70% of the prioritizations. To allow comparison between data sources we plotted rank receiver operating characteristic (ROC) curves, from which sensitivity/specificity values can be easily deduced. The area under this curve (AUC) is a standard measure of the performance of this algorithm. For instance, an AUC-value of 100% indicates that every defector gene ranked first, whereas a value of 50% means that the defector genes ranked randomly.

For every single data source, Endeavour reached a higher AUC score for disease and pathway genes than for randomly selected genes, indicating that it was sensitive and specific in ranking the defector gene, regardless of the type of data source consulted (**Fig. 2**). Not surprisingly, the data sources differed in their usefulness and suitability to rank genes (**Supplementary Notes**).

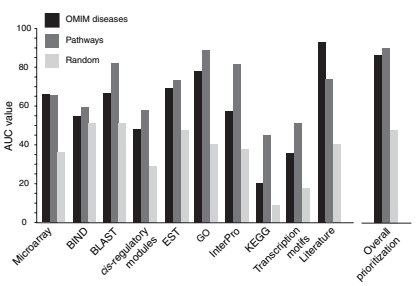

Figure 2 Cross-validation results. The AUC values obtained for all individual data sources are shown for disease prioritizations (black), pathway prioritizations (dark gray) and random prioritizations (light gray). The AUC values from the overall prioritization obtained after fusing all individual prioritizations are also shown.

Overall prioritization by fusing multiple data sources

Although in most cases the defector gene ranked high in the prioritization list, this was not always the case (**Supplementary Fig. 1** online).

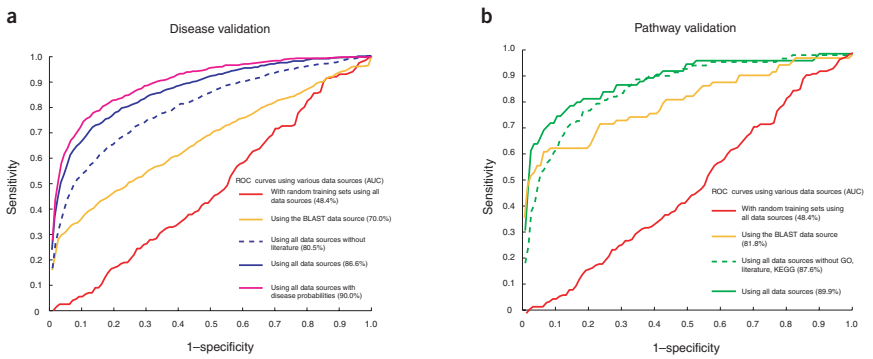

Figure 3 Cross-validation results. (**a**) Rank ROC curves obtained for the disease validation. (**b**) Rank ROC curves obtained for the pathway validation. In both figures, the control ROC curve (red line) was obtained after prioritization with randomly constructed training sets and by using all data sources. For all other ROC curves, disease or pathway-specific training sets were generated. The data sources used to construct every ROC curve are indicated on the figure.

To minimize this variability and to increase the performance of ranking, we integrated all individual prioritizations into a single overall rank by implementing an algorithm based on order statistics. With this algorithm, the probability of finding a gene at all the observed positions is calculated and a single overall rank is obtained by ranking genes according to these probabilities. To evaluate the performance of this overall ranking, we calculated its AUC values, as described above for the individual data sources. The AUC scores were 86.6% and 89.9% for disease and pathway genes compared to 48.4% for randomly selected genes (**Fig. 3a**,**b**). The correct pathway gene ranked among the top 50% of test genes in 95% of the cases, or among the top 10% in 74% of the cases. The variability of the overall prioritization was substantially smaller than that of individual data sources (**Supplementary Fig. 1**), and each of the data sources contributed to the overall ranking (**Supplementary Fig. 2** online). Our validation experiment thus results in biologically meaningful prioritizations.

Almost every data source but especially functionally annotated databases are incompletely annotated. For instance, only 63% of the genes are currently annotated in the GO database. Consequently, existing methods using these data sources introduce an undesired bias toward better-studied genes. Our approach should suffer less from these shortcomings as it also uses sequence-based sources containing information about known and unknown genes. In support of this, we found that the overall ranking of defector genes was not substantially influenced by the number of data sources if at least three sources with data annotations were available (**Supplementary Fig. 3a** online). In fact, even unknown genes lacking a

Table 1 Prioritizations of recently identified monogenic disease genes

				Rank position using the indicated data sources	
Disease	Gene	Ensembl ID	Publication date	All	Literature
Arrhythmia	*CACNA1C*	ENSG00000151067	October 2004 (ref. 34)	4	3
Congenital heart disease	*CRELD1*	ENSG00000163703	April 2003 (ref. 35)	3	1
Cardiomyopathy 1	*CAV3*	ENSG00000182533	January 2004 (ref. 36)	2	1
Parkinson disease	*LRRK2*	ENSG00000188906	November 2004 (ref. 37)	50	*
Charcot-Marie-Tooth disease	*DNM2*	ENSG00000079805	March 2005 (ref. 38)	14	100
Amyotrophic lateral sclerosis	*DCTN1*	ENSG00000135406	August 2004 (ref. 39)	27	97
Klippel-Trenaunay disease	*AGGF1* (also known as *VG5Q*)	ENSG00000164252	February 2004 (ref. 40)	3	39
Cardiomyopathy 2	*ABCC9*	ENSG00000069431	April 2004 (ref. 41)	1	51
Distal hereditary motor neuropathy	*BSCL2*	ENSG00000168000	March 2004 (ref. 42)	15	62
Cornelia de Lange syndrome	*NIPBL*	ENSG00000164190	June 2004 (refs. 43,44)	9	75
			Average rank	13 ± 5	48 ± 13

For all genes, a mutation was inherited in a mendelian fashion (or was shown to cause the disease phenotype). The name of the disease and disease-causing gene, the Ensembl ID and the publication date of the article reporting the gene mutation (month-year) are shown, together with the rank (out of 200 test genes) at which they were prioritized by Endeavour, using all data sources or using the pre-publication date literature source alone. The average rank (mean ± s.e.m.) for each prioritization is indicated. For *LRRK2*, no literature information was available. This has been indicated in the table by an asterisk (*).

Table 2 Prioritizations of recently identified polygenic disease genes

Disease	Gene	Ensembl ID	Publication date	Rank
Atherosclerosis 1	*TNFSF4*	ENSG00000117586	April 2005 (ref. 45)	54
Crohn disease	*SLC22A4, SLC22A5*	ENSG00000197208	May 2004 (ref. 46)	71
Parkinson disease	*GBA*	ENSG00000188906	November 2004 (47)	23
Rheumatoid arthritis	*PTPN22*	ENSG00000134242	August 2004 (ref. 48)	11
Atherosclerosis 2	*ALOX5AP*	ENSG00000132965	February 2004 (ref. 49)	29
Alzheimer disease	*UBQLN1*	ENSG00000135018	March 2005 (ref. 50)	54
			Average rank	40 ± 10

The nature of the genetic variation in these genes was in each case a polymorphism, which typically was inherited as a risk factor for the respective disease. The name of the complex disease in which these genes were identified, their gene name, Ensembl ID and the publication date when the disease gene was reported as a susceptibility gene are given, together with the rank (out of 200 test genes) at which they have been prioritized by all data sources with rolled-back literature. The relative contribution of these genetic variations as risk factors for disease susceptibility will become clearer once replication studies are performed. The average rank (mean ± s.e.m.) for each prioritization is indicated.

HUGO name and with very little information available could be ranked highly (**Supplementary Fig. 3b**). Thus, our method takes into account data sources with relevant information, while disregarding noninformative ones. This may be particularly advantageous for the prioritization of disease genes, as unknown genes are not readily considered as disease candidates when selected manually.

Endeavour does not rely on literature-derived data alone

For each OMIM gene used in the disease validation, a mutation causing the disease had previously been reported in a landmark study. Because the inclusion of these publications may artificially increase the relative contribution of the literature data source in the overall performance of this algorithm, we excluded, as a test, the entire literature database from the disease validation protocol. For the same reason, the GO, KEGG and literature data sources were excluded from the pathway validation. Even under such unrealistic conditions where entire data sources were not used, the overall performance of the algorithm was only negligibly affected: the performance dropped by only 6.1% for disease genes (from 86.6% to 80.5%; **Fig. 3a**) and by only 2.3% for pathway genes (from 89.9% to 87.6%; **Fig. 3b**). Thus, the diversity of data sources used in our approach enables meaningful prioritizations, even without the use of literature information.

Clearly, this caution is only of importance in the context of a validation. In a more realistic situation, when the precise function of a disease gene is not known yet, the literature could still provide valuable indirect information about other properties of a gene. In a study of ten monogenic diseases (see below), we mimicked this situation by using only 'rolled-back' literature information, available one year before the landmark publication. Even then Endeavour provided a high rank for three genes (position 1, 1 and 3 out of 200 test genes, **Table 1**), illustrating that the literature contributes to the prioritization of yet undiscovered disease genes. For the seven other genes, use of the literature as the only data source was not very efficient, but inclusion of all the other data sources yielded a high rank (**Table 1**). Overall, even though the literature may provide valuable information, our method does not rely on literature as the only critical data source. But also, its performance is not restricted by the lack of available literature data, because of its ability to access and integrate multiple other data sources.

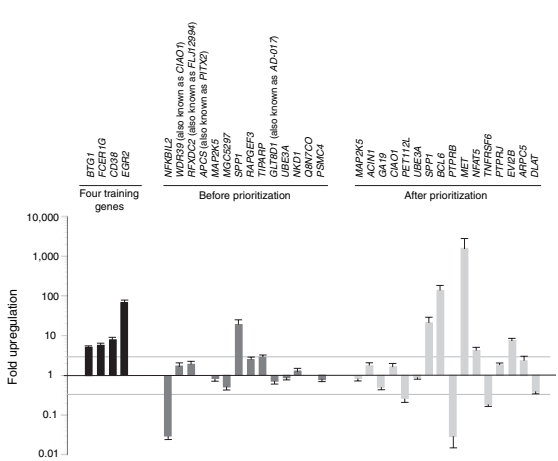

Figure 4 *In vitro* functional validation of Endeavour. Results of real-time quantitative PCR measurements in differentiated versus undifferentiated HL-60 cells. Expression profiles of 4 out of 18 training genes (left), which were tested as a positive control, and 20 target genes predicted by the *cis*-regulatory module model (center) are shown. Expression levels of *SPP1* and *NGKBIL2* differed more than threefold between differentiated and undifferentiated cells; expression levels for six genes could not be measured. The expression profiles of the 20 highest-ranking target genes after prioritization by Endeavour (right) are also shown. Expression levels of eight genes (*SPP1, BCL6, PTPRB, MET, TNFRSF6, NFAT5, PET112L* and *EVI2B*) differed more than threefold between differentiated and undifferentiated cells; four genes could not be measured. The fold difference is depicted on a logarithmic scale; error bars represent the s.e.m. The line indicates the threshold (threefold up- or downregulation).

Use of disease-specific data sources

An important asset of Endeavour is that its framework was designed to allow the inclusion of additional data sources, such as disease-related features, in the prioritization strategy. We illustrate this for the prioritization of disease genes. On the basis of a number of selection criteria (e.g., protein length, phylogenetic conservation), Lopez-Bigas and Adie determined for every gene a 'general' disease probability, or its probability as a disease candidate gene[8,9]. When integrating the Lopez-Bigas or Adie criteria in Endeavour as an additional data source, we found that its performance improved further (AUC scores increased by up to 5% regardless of the inclusion of literature sources). Likewise, microarray data specific for the process or disease under study can be included. Our approach thus allows the user to add, in a flexible and modular manner, additional data sources, such as appropriate disease-specific data sources, to enhance its overall performance.

Prioritization of genes causing monogenic diseases

In the large-scale validation, 627 genes were automatically selected from the OMIM database, without taking their mono- or polygenic nature into account. We therefore assessed whether our approach could be used to prioritize genes that cause monogenic diseases. As experimentalists often prefer to select their own sets of training genes, instead of relying on automatically derived genes or characteristics, we selected ten monogenic diseases and constructed sets of training genes together with a biological expert (**Table 1** and **Supplementary Table 1**). To simulate the real life situation, we deliberately chose recently identified disease-causing genes, and used rolled-back literature together with all other data sources. The set of test genes included the gene causing the monogenic disease, and 199 genes flanking its immediate chromosomal surroundings. The algorithm gave the ten monogenic disease–causing genes an average rank of 13 ± 5 out of 200 test genes (**Table 1**). When using a training set not related to the disease under study to prioritize the test sets as a negative control, the disease genes ranked randomly (position 96 on average). As a further validation the algorithm was applied to a very large set of test genes (that is, all 1,048 genes from chromosome 3; **Supplementary Notes** and **Supplementary Table 2** online).

This pseudo-prospective analysis, using rolled-back literature, reveals that expert-based construction of training sets may lead to high discovery rates when hunting for monogenic disease genes in both small and large test sets.

Prioritization of genes underlying polygenic diseases

In many cases, human disease is not monogenic, but polygenic in nature. We therefore prioritized six genes, recently identified as polygenic disease genes, together with 199 chromosomal flanking genes (**Table 2**). The sets of training genes used for these prioritizations are explained in **Supplementary Table 1**. On average, the susceptibility genes ranked at position 40 ± 10, when using the rolled-back literature together with all the other data sources. As expected, the prioritization of polygenic disease candidate genes is a greater challenge than ranking monogenic disease genes. Nonetheless, the ranking was still specific, as the susceptibility genes ranked at position 96 ± 10, when training sets for these disorders were randomly assigned to other test sets as a negative control. Thus, although the performance is lower than for monogenic diseases (as anticipated), susceptibility genes to polygenic diseases can be enriched by Endeavour's prioritization.

Prioritization of regulatory pathway genes

To analyze whether Endeavour could also rank genes involved in a particular biological process, we combined computation with functional validation *in vitro*. First, using the previously characterized ModuleSearcher

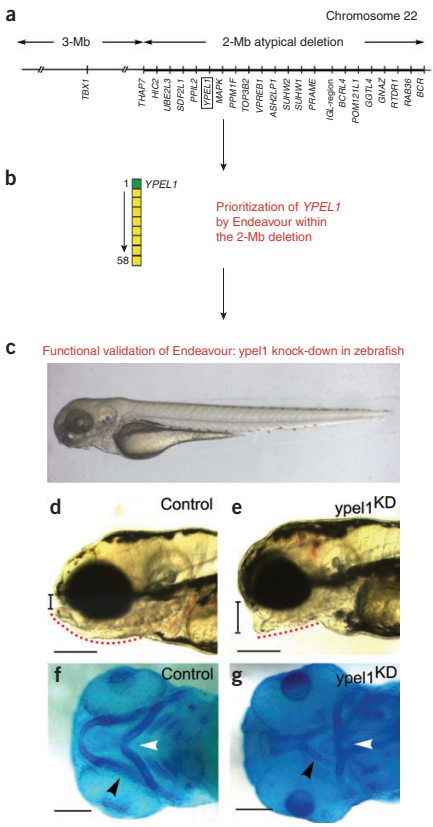

Figure 5 Functional validation of Endeavour in zebrafish. (**a**) Part of chromosome 22, illustrating the hemizygous 3-Mb region deleted in many DGS patients and the atypical 2-Mb region, which is deleted in some (atypical) DGS patients. For clarity, only some of the 58 Ensembl-annotated genes within the 2-Mb region, and only *TBX1* in the 3-Mb deleted region, are shown. It remains unknown whether any of the genes in the 2-Mb region play a role in pharyngeal arch development defects seen in DGS. (**b**) *YPEL1* was prioritized among the 58 genes of the 2-Mb deleted region by Endeavour as the most likely candidate involved in pharyngeal arch development. (**c**) Photo of a zebrafish, which has been used as a suitable model to study the role of *YPEL1* in pharyngeal arch development. (**d**,**e**) Lateral view of the head in live embryos at 4 d after fertilization. The lower jaw is clearly visible in the control, whereas ypel1^{KD} embryos show an underdeveloped lower jaw (mandibular arch; indicated by the red dotted line) and open mouth (indicated by the vertical line). (**f**,**g**) Ventral view of the pharyngeal arch cartilage using alcian blue stain at 3 d after fertilization. Black arrow depicts the mandibular arch; white arrow depicts hyoid arch. In ypel1^{KD} embryos, the jaw arches were severely malformed with the mandibular arch often reduced in size. The pharyngeal arch cartilage also showed reduced or no staining.

Table 3 Prioritization of *YPEL1* by Endeavour

Training sets used to prioritize *TBX1* or *YPEL1*	Rank assigned to *YPEL1*	Rank assigned to *TBX1*
DGS-related		
DGS (14)	1*	1*
Cardiovascular birth defects (14)	3*	1*
Cleft palate birth defects (9)	2*	1*
Neural crest genes (14)	1*	2*
Average rank	1.75 ± 0.48	1.25 ± 0.25
DGS-unrelated		
Atherosclerosis (24)	12	24
Parkinson disease (9)	31	15
Distal hereditary motoneuropathy (8)	13	41
Charcot-Marie-Tooth disease (17)	9	16
Alzheimer's disease (5)	21	14
Rheumatoid arthritis (8)	20	7
Inflammatory bowel disease (7)	7	24
Average rank	16 ± 3	20 ± 4

The set of test genes contained the 58 genes present in the 2-Mb atypical deletion region on chromosome 22q11 (middle column) or, in addition, the *TBX1* gene (right column). These test genes were prioritized by Endeavour for their similarity to the indicated set of training genes, which were related or unrelated to DGS. As shown, *TBX1* and *YPEL1* ranked among the first three test genes, indicating their high degree of similarity with the set of training genes (*, probability of $P < 0.05$ that the test and training genes had a similar profile). The number of training genes is indicated between brackets.

algorithm within TOUCAN[12,13], we predicted a *cis*-regulatory module (CRM) in the regulatory regions of 18 genes, known to be upregulated during myeloid differentiation[14]. We then selected 100 putative target genes containing this CRM from the genome, and ordered them according to their CRM score (see **Supplementary Notes**). These 100 genes were then prioritized with the algorithm, using the 18 genes involved in myelopoiesis as a training set. To investigate whether it enriched the number of true-positive target genes involved in myeloid differentiation, we induced differentiation of HL-60 cells *in vitro* and analyzed which of the 20 best ranking genes, before and after prioritization by Endeavour, were more than threefold up- or downregulated. Before prioritization, the expression of two genes (**Fig. 4**) was differentially regulated, whereas after prioritization up to eight genes were differentially regulated ($P < 0.05$; **Fig. 4**). Importantly, several of these differentially regulated genes are implicated in myeloid function: *SPP1*, *BCL6* and *MET* are known to be involved in myeloid differentiation[15–17], whereas *FRSF6*, better known as the *FAS* inducer of apoptosis, is a suppressor of macrophage activation[18]. The possible involvement of *PTPRB*, *NFAT5*, *PET112L* and *EVI2B* in myeloid differentiation was, however, unknown. Our prioritization protocol can thus be used for gene discovery as well.

Functional validation of Endeavour in zebrafish

As a final and most stringent test, we validated our approach in an animal model *in vivo*. The DiGeorge syndrome (DGS) is a common congenital disorder, in which craniofacial dysmorphism and other defects result from abnormal development of the pharyngeal arches[19,20]. Many DGS patients typically have a 3-Mb hemizygous deletion in chromosome 22 (*del22q11*)[19,20]. Genetic studies in mice and zebrafish have established *Tbx1* as a key DGS disease candidate gene in this region[21–24] (**Fig. 5a**). In atypical DGS cases, a 2-Mb region, downstream of *del22q11* is deleted[25], but it remains unknown which of the 58 Ensembl-annotated genes in this region plays a role in pharyngeal arch development. In this experiment, we first assessed whether the algorithm would prioritize any of these genes as a possible regulator of pharyngeal arch development, and then analyzed whether this gene indeed affected this process *in vivo*.

We first tested, as a positive control, whether Endeavour would identify *TBX1* as a DGS candidate when added to the list of 58 test genes. To avoid possible selection bias due to an overly restricted choice of training genes, we used various training sets according to their relationship with DGS, cardiovascular or cleft palate birth defects (typical DGS symptoms), or neural crest biology (neural crest cell anomalies cause DGS-like symptoms; **Supplementary Notes**). When using these training sets, *TBX1* ranked first or second (**Table 3**). This prioritization was specific, as *TBX1* was not identified as a DGS candidate gene when using training genes unrelated to DGS. We then used our approach to prioritize the 58 genes of the 2-Mb deleted region. When using various sets of DGS-related training genes, the top-ranking gene was always *YPEL1* (**Table 3** and **Fig. 5b**). Similar to the *TBX1* simulation, use of a set of training genes, unrelated to DGS, confirmed that the prioritization was specific for DGS.

To assess the functional role of *YPEL1 in vivo*, we used the zebrafish model, which has been previously used as a suitable model to study pharyngeal arch development[26] (**Fig. 5c**). Ypel1 protein levels in zebrafish embryos were knocked down using a set of antisense morpholino oligonucleotides (morpholinos), each targeting different sequences of the *ypel1* transcript and dose-dependently and specifically inhibiting *ypel1* translation (not shown). The role of *ypel1* in pharyngeal arch morphogenesis was evaluated by phenotyping the development of its derivatives, that is, the jaws and other skeletal structures of the skull[27]. Ypel1 knockdown (ypel1^{KD}) embryos displayed various craniofacial defects. In particular, they exhibited an underdeveloped jaw, with the most severely affected embryos displaying an open-mouth phenotype suggestive of craniofacial dysmorphism (**Fig. 5d**,**e**). Ypel1^{KD} embryos also displayed defects in pharyngeal arch cartilage formation, ranging from an overall disorganization to a complete loss of the jaw and pharyngeal arch cartilage. In some ypel1^{KD} embryos, the mandibular arch was strongly reduced in size. Occasionally, no staining of cartilage could be detected at all (**Fig. 5f**,**g**). Ypel1^{KD} embryos exhibited additional pharyngeal arch defects, which will be described in more detail elsewhere.

In summary, our method identified *YPEL1* as a candidate DGS gene and *in vivo* experiments confirmed its role in pharyngeal arch development. These data raise the intriguing question whether *YPEL1* might be a novel disease candidate gene of atypical DGS in humans.

DISCUSSION

The number of publicly available databases containing information about human genes and proteins continues to grow. Here, we developed a method to integrate all this information and prioritize any set of genes based on their similarity to a set of reference genes. Such a prioritization is not only useful for gene hunting in human diseases, but also for identifying members of biological processes.

Our approach is useful in several respects. First, it uses genes to retrieve information about a disease or biological pathway, instead of disease characteristics. Existing methods that use disease characteristics can only retrieve information from databases that use the same disease vocabulary[4,5,7]. By using genes, Endeavour accesses not only these vocabulary-based data sources, but also other data sources, storing

information about a gene (e.g., derived from a microarray experiment) or a gene sequence (e.g., BLAST sequence similarity). Moreover, by using genes, the method is also suitable for gene prioritization in biological processes as well.

Second, compared to existing methods, which access only one or two data sources[4–7], our method accesses many more data sources (currently up to 12). Importantly, consultation of each of the individual sources by Endeavour generates biologically relevant prioritizations. We developed an algorithm based on order statistics to fuse all these separate prioritizations into a single overall rank. This algorithm is able to handle genes with missing values, thereby minimizing the bias for known or well-characterized genes. This bias will decrease even further in the future, when new and better high-throughput data become available, and when the genome annotation and curation processes reach their finalization.

Third, the algorithm is publicly available as a software tool, built by bioinformaticians, but designed for experimentalists, helping them to focus readily on key biological questions. The only other available prioritization tool for diseases, G2D, uses GO and literature data sources and is therefore restricted in making predictions about annotated or known genes[5].

Fourth, the approach gives the user maximal control over the set of training and test genes. Biologists prefer the flexibility of interactively selecting their own set of genes over an automatic and noninteractive data-mining selection procedure.

We validated the method extensively, in a large-scale validation study of 703 disease and pathway genes, and in a number of case-specific analyses. The validation results were remarkably good: on average, the correct gene was ranked 10th out of 100 test genes—for monogenic diseases, the performance was even better. The algorithm was capable of prioritizing large test sets (up to 1,000 genes)—the upgrade of Endeavour into a package capable of prioritizing the entire genome would be an interesting perspective for the future. Functional validation studies *in vitro* further demonstrated that the method worked equally well for prioritization of pathway genes. Furthermore, *in vivo* studies in zebrafish revealed that *YPEL1*, a gene prioritized by Endeavour in a 2-Mb chromosomal region deleted in patients with craniofacial defects, indeed regulates morphogenesis of the pharyngeal arches and their craniofacial-derivative structures.

Lastly, the Endeavour software design is modular and allows the inclusion of publicly available or proprietary data sources (e.g., disease-specific microarray experiments). We have illustrated and validated this possibility by including the general disease probability criteria of Lopez-Bigas[9] and Adie[8].

In summary, we present a computational method for fast and interactive gene prioritization that fuses genomic data regardless of its origin.

METHODS

Data sources. A more detailed description of the data sources is available as **Supplementary Methods** online. Briefly, for information retrieved from attribute-based data sources (that is, Gene Ontology, EST expression, InterPro and KEGG), the algorithm uses a binomial statistic to select those attributes that are statistically overrepresented among the training genes, relative to their genome-wide occurrence. Each overrepresented attribute receives a *P*-value p_i that is corrected for multiple testing. For information retrieved from vector-based data sources (that is, literature, microarray expression data or *cis*-regulatory motif predictions), the algorithm constructs an average vector profile of the training set. The literature profile is based on indexed abstracts and contains inverse document frequencies for each term of a GO-based vocabulary[28]; the expression profile contains expression ratios; the motif profile contains scores of TRANSFAC position weight matrices, obtained by scanning promoter sequences of the training genes that are conserved with their respective mouse orthologous sequences. To rank a set of test genes, attribute-based data are scored by Fisher's omnibus meta-analysis (Σ-2logp_i), generating a new *P*-value from a χ^2-distribution. Vector-based data are scored by Pearson correlation between the test vector and the training average. The data in the BLAST, BIND and *cis*-regulatory module (CRM) databases are neither vector- nor attribute-based. For BLAST, the similarity score between a test gene and the training set is the lowest *e*-value obtained from a BLAST against an *ad hoc* indexed database consisting of the protein sequences of the training genes. For BIND (Biomolecular Interaction Network Database)[29], the similarity score is calculated as the overlap between all protein-protein interaction partners of the training set and those of the test gene. Lastly, for CRM data, the best combination of five clustered transcription factor binding sites—in all human-mouse conserved noncoding sequences (up to 10 kb upstream of transcription start site) of the training genes—is determined using a genetic algorithm[12,30]. The similarity of this trained model to a test gene is determined by scoring this motif combination on the conserved noncoding sequences of the test gene.

Order statistics. The rankings from the separate data sources are combined using order statistics. A *Q* statistic is calculated from all rank ratios using the joint cumulative distribution of an *N*-dimensional order statistic as previously done by Stuart *et al.*[31]

$$Q(r_1,r_2,\ldots,r_N) = N!\int_0^{r_1}\int_{s_1}^{r_2}\ldots\int_{s_{N-1}}^{r_N} ds_N ds_{N-1}\ldots ds_1 \qquad (1)$$

They propose the following recursive formula to compute the above integral:

$$Q(r_1,r_2,\ldots,r_N) = N!\sum_{i=1}^{N}(r_{N-i+1}-r_{N-i})Q(r_1,r_2,\ldots,r_{N-i},r_{N-i+2},\ldots,r_N) \qquad (2)$$

where r_i is the rank ratio for data source i, N is the number of data sources used, and $r_0 = 0$. However, two problems arose when we tried to use this formula. First, we noticed that this formula is highly inefficient for moderate values of N, and even intractable for $N > 12$ because its complexity is $O(N!)$. We therefore implemented a much faster alternative formula with complexity $O(N^2)$:

$$V_k = \sum_{i=1}^{k}(-1)^{i-1}\frac{V_{k-i}}{i!}r_{N-k+1}^{i} \qquad (3)$$

with $Q(r_1,r_2,\ldots,r_N) = N!V_N$, $V_0 = 1$, and r_i is the rank ratio for data source i.

Second, we noticed that the *Q* statistics calculated by (1) are not uniformly distributed under the null hypothesis and can thus not directly be used as *P*-values. Therefore, we fitted a distribution for every possible number of ranks and used this distribution to calculate an approximate *P*-value. We found that the *Q* statistics for $N \leq 5$ randomly and uniformly drawn rank-ratios are approximately distributed according to a beta distribution. For $N > 5$ the distributions can be modeled by a gamma distribution. The cumulative distribution function of these distributions provides us with a *P*-value for every *Q* statistic from the order statistics. Next to the original *N* rankings, we now have an $(N + 1)^{th}$ that is the combined rank of all separate ranks.

Cell culture, RNA isolation and RT-PCR. HL-60 cells were grown in RPMI 1640, supplemented with 10% FCS. Differentiation was induced by 10 nM phorbol 12-myristate 13-acetate (PMA), when cells were grown to a density of 7×10^5/ml. Before induction and 24 h after induction, cells were harvested by centrifugation and RNA was isolated using the trizol reagent (Invitrogen), and subsequently treated with Turbo DNA-free DNase (Ambion). First-strand cDNA was synthesized using Superscript II reverse transcriptase (Invitrogen). Real-time quantitative PCR was performed using the qPCR core kit for SYBR green (Eurogentec), on an ABI PRISM 7700 SDS (Applied BioSystems). The mRNA levels were normalized to the geometric average of four different housekeeping genes: *ACTB*, *GAPDH*, *UBC* and *HPRT1*. Numbers of differentially expressed genes before and after prioritization were compared with a chi-square test.

Zebrafish care and embryo manipulations. Wild-type zebrafish (*Danio rerio*) of the AB strain were maintained under standard laboratory conditions[32]. Morpholino oligonucleotides (Gene Tools) were injected into one- to four-cell-stage embryos[27]. Alcian blue cartilage staining was carried out as previously described[33]. All animal studies were reviewed and approved by the institutional animal care and use committee for Medical Ethics and Clinical Research of the University of Leuven.

Software availability. Endeavour is freely available for academic use as a Java application at http://www.esat.kuleuven.be/endeavour.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We wish to thank all groups and consortia that made their data freely available: Ensembl, NCBI (EntrezGene and Medline), Gene Ontology, BIND, KEGG, Atlas, InterPro, BioBase, the Disease Probabilities from Lopez-Bigas and Ouzounis[9] and the Prospectr scores from Euan Adie[8]. Ouzounis[8] and the Prospectr scores from Euan Adie[9]. We also thank the following people for their help in particular areas: Robert Vlietinck with the manuscript, Patrick Glenisson with text mining, Joke Allemeersch and Gert Thijs with the order statistics and Camilla Esguerra with the zebrafish experiments. S.A., D.L. and P.V.L. are sponsored by the Research Foundation Flanders (FWO). This work is supported by Flanders Institute for Biotechnology (VIB), Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (IWT) (STWW-00162), Research Council KULeuven (GOA-Ambiorics, IDO genetic networks), FWO (G.0229.03 and G.0413.03), IUAP V-22, K.U.L. Excellentiefinanciering CoE SymBioSys (EF/05/007), EU NoE Biopattern and EU EST BIOPTRAIN to Y.M., and by the FWO (G.0405.06), GOA/2006/11 and GOA/2001/09, Squibb and EULSHB-CT-2004-503573 to P.C.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Quackenbush, J. Genomics. Microarrays—guilt by association. *Science* **302**, 240–241 (2004).
2 Kanehisa, M. & Bork, P. Bioinformatics in the post-sequence era. *Nat. Genet.* **33** Suppl. 305–310 (2003).
3. Ball, C.A., Sherlock, G. & Brazma, A. Funding high-throughput data sharing. *Nat. Biotechnol.* **22**, 1179–1183 (2004).
4. Freudenberg, J. & Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18** Suppl. 2, S110–S115 (2002).
5. Perez-Iratxeta, C., Bork, P. & Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* **31**, 316–319 (2002).
6. Turner, F.S., Clutterbuck, D.R. & Semple, C.A. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.* **4**, R75 (2003).
7. Tiffin, N. *et al.* Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.* **33**, 1544–1552 (2005).
8. Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. & Pickard, B.S. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6**, 55 (2005).
9. Lopez-Bigas, N. & Ouzounis, C.A. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* **32**, 3108–3114 (2004).
10. Kent, W.J. *et al.* Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* **15**, 737–741 (2005).
11. Altermann, E. & Klaenhammer, T.R. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics* **6**, 60 (2005).
12. Aerts, S. *et al.* TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.* **33**, W393–W396 (2005).
13. Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. & De Moor, B. Computational detection of cis-regulatory modules. *Bioinformatics* **19** (Suppl 2), II5–II14 (2003).
14. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).
15. Stegmaier, K. *et al.* Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nat. Genet.* **36**, 257–263 (2004).
16. Pixley, F.J. *et al.* BCL6 suppresses RhoA activity to alter macrophage morphology and motility. *J. Cell Sci.* **118**, 1873–1883 (2005).
17. Galimi, F. *et al.* Hepatocyte growth factor is a regulator of monocyte-macrophage function. *J. Immunol.* **166**, 1241–1247 (2001).
18. Brown, N.J. *et al.* Fas death receptor signaling represses monocyte numbers and macrophage activation in vivo. *J. Immunol.* **173**, 7584–7593 (2004).
19. Scambler, P.J. The 22q11 deletion syndromes. *Hum. Mol. Genet.* **9**, 2421–2426 (2000).
20. Baldini, A. Dissecting contiguous gene defects: TBX1. *Curr. Opin. Genet. Dev.* **15**, 279–284 (2005).
21. Jerome, L.A. & Papaioannou, V.E. DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. *Nat. Genet.* **27**, 286–291 (2001).
22. Merscher, S. *et al.* TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome. *Cell* **104**, 619–629 (2001).
23. Lindsay, E.A. *et al.* Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature* **410**, 97–101 (2001).
24. Piotrowski, T. *et al.* The zebrafish van gogh mutation disrupts tbx1, which is involved in the DiGeorge deletion syndrome in humans. *Development* **130**, 5043–5052 (2003).
25. Rauch, A. *et al.* A novel 22q11.2 microdeletion in DiGeorge syndrome. *Am. J. Hum. Genet.* **64**, 659–666 (1999).
26. Graham, A. The development and evolution of the pharyngeal arches. *J. Anat.* **199**, 133–141 (2001).
27. Stalmans, I. *et al.* VEGF: a modifier of the del22q11 (DiGeorge) syndrome? *Nat. Med.* **9**, 173–182 (2003).
28. Glenisson, P. *et al.* TXTGate: profiling gene groups with text-based information. *Genome Biol.* **5**, R43 (2004).
29. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
30. Aerts, S., Van Loo, P., Moreau, Y. & De Moor, B. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics* **20**, 1974–1976 (2004).
31. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
32. Westerfield, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish*, (University of Oregon Press, Eugene, Oregon, 1994).
33. Kimmel, C.B. *et al.* The shaping of pharyngeal cartilages during early development of the zebrafish. *Dev. Biol.* **203**, 245–263 (1998).
34. Splawski, I. *et al.* Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**, 19–31 (2004).
35. Robinson, S.W. *et al.* Missense mutations in CRELD1 are associated with cardiac atrioventricular septal defects. *Am. J. Hum. Genet.* **72**, 1047–1052 (2003).
36. Hayashi, T. *et al.* Identification and functional analysis of a caveolin-3 mutation associated with familial hypertrophic cardiomyopathy. *Biochem. Biophys. Res. Commun.* **313**, 178–184 (2004).
37. Zimprich, A. *et al.* Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **44**, 601–607 (2004).
38. Zuchner, S. *et al.* Mutations in the pleckstrin homology domain of dynamin 2 cause dominant intermediate Charcot-Marie-Tooth disease. *Nat. Genet.* **37**, 289–294 (2005).
39. Munch, C. *et al.* Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. *Neurology* **63**, 724–726 (2004).
40. Tian, X.L. *et al.* Identification of an angiogenic factor that when mutated causes susceptibility to Klippel-Trenaunay syndrome. *Nature* **427**, 640–645 (2004).
41. Bienengraeber, M. *et al.* ABCC9 mutations identified in human dilated cardiomyopathy disrupt catalytic KATP channel gating. *Nat. Genet.* **36**, 382–387 (2004).
42. Windpassinger, C. *et al.* Heterozygous missense mutations in BSCL2 are associated with distal hereditary motor neuropathy and Silver syndrome. *Nat. Genet.* **36**, 271–276 (2004).
43. Tonkin, E.T., Wang, T.J., Lisgo, S., Bamshad, M.J. & Strachan, T. NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. *Nat. Genet.* **36**, 636–641 (2004).
44. Krantz, I.D. *et al.* Exclusion of linkage to the CDL1 gene region on chromosome 3q26.3 in some familial cases of Cornelia de Lange syndrome. *Am. J. Med. Genet.* **101**, 120–129 (2001).
45. Wang, X. *et al.* Positional identification of TNFSF4, encoding OX40 ligand, as a gene that influences atherosclerosis susceptibility. *Nat. Genet.* **37**, 365–372 (2005).
46. Peltekova, V.D. *et al.* Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat. Genet.* **36**, 471–475 (2004).
47. Aharon-Peretz, J., Rosenbaum, H. & Gershoni-Baruch, R. Mutations in the glucocerebrosidase gene and Parkinson's disease in Ashkenazi Jews. *N. Engl. J. Med.* **351**, 1972–1977 (2004).
48. Begovich, A.B. *et al.* A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **75**, 330–337 (2004).
49. Helgadottir, A. *et al.* The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat. Genet.* **36**, 233–239 (2004).
50. Bertram, L. *et al.* Family-based association between Alzheimer's disease and variants in UBQLN1. *N. Engl. J. Med.* **352**, 884–894 (2005).

3.2 Contribution of the PhD candidate

The PhD candidate has contributed to the programing of the Endeavour software that implements the gene prioritization strategy described in the publication. More precisely, he has programed part of the Java interface that is used to load or save the gene sets to be used in the analysis. In addition, he has also built up a semi automated update system for the human genomic data sources.

3.3 Discussion

The approach developed in the present paper relies on the 'guilt-by-association' concept so that the predicted disease genes are the in fact similar to the ones already known to be involved in the disease under study. It means that predictions are optimal when the underlying data is suggesting that the hidden disease genes could be related, even indirectly, with the training genes and that therefore the associations can be predicted. A corollary is that it will not work when the hidden disease gene association is not backed-up by any genomic data, meaning that it is impossible to predict associations that are coming *ex nihilo*. This needs to be kept in mind when analyzing validation results, because some of the associations can simply not be predicted given the data.

A characteristic of our approach is the use of a set of genes to model the biological process of interest. This allows a fine tuning of the training since genes can be added or removed individually (as compared to when the disease name is used for training, which does not leave a lot of space for fine tuning). However, this can also be a disadvantage when only a few or even no disease genes are found. A proposed approach is to use the genes involved in the biomolecular pathways that are disturbed in the disease process. This can however still be a challenging task since sometimes no pathway seems to be critical or, at the contrary, many pathways are concurrently disturbed. Furthermore, the building of a training set is a time consuming process that requires some expertise, which does not facilitate the use of the tool. This issue is further discussed in chapter 9.

Since the publication of that paper in 2006, the existing data sources have been updated several times to reflect more accurately the current knowledge in human biology and genetics. Furthermore, several additional data sources have been integrated into the workflow. The performance of Endeavour on the same benchmark (29 genetic diseases and 3 signaling pathways) has increased correspondingly as can be shown in figure 3.1. The algorithm and the benchmark gene sets are the same so the observed differences are only coming from the data sources (the data sources have been updated, but the data sources that have been added have not been used). The gain is significant (p-value < 0.05, Wilcoxon

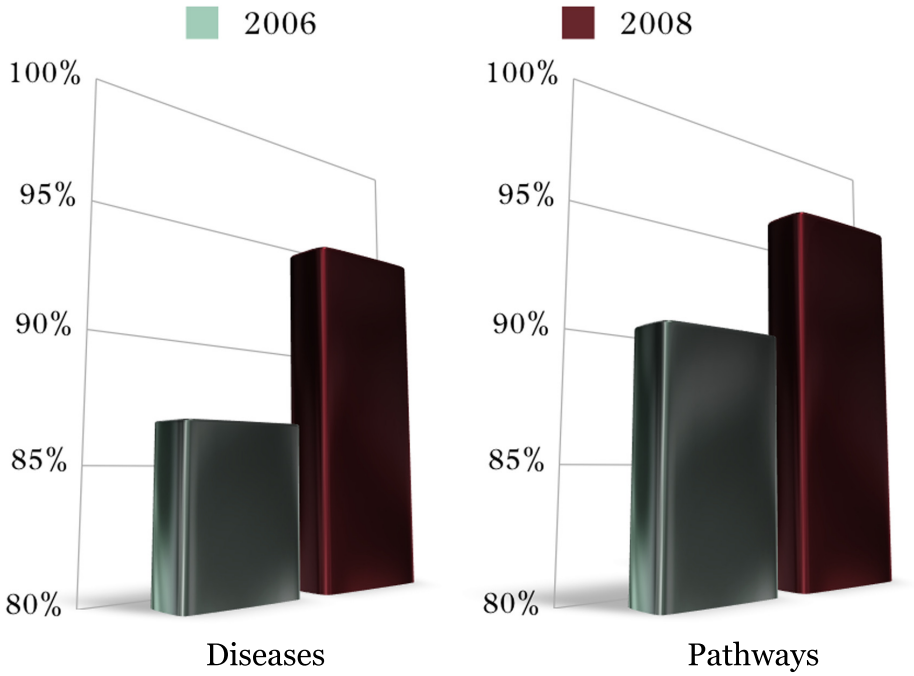

Figure 3.1: Comparison of the performance on the same benchmark at different time points. The AUC is plotted in the y-axis for the 29 genomic diseases (620 genes) on the left and for the 3 signaling pathways (75 genes) on the right. The AUC is plotted for the genomic data sources of 2006 (grey - left columns) and of 2008 (red - right columns). The difference is significant for the genomic diseases and for the signaling pathways (p-value < 0.05, Wilcoxon rank-sum).

rank-sum) and indicates that as the databases grow, our representation of the biological knowledge gets better and the predictions become more accurate.

The strategy developed in this publication has also been benchmarked through an extensive benchmark based on MetaCore disease marker sets and pathway maps. This and other benchmark experiments are described in chapter 7. The method has also been extended to cover model organisms as described in chapter 4, and was further experimentally validated as described in chapter 8. In addition, a second gene prioritization strategy was developed and is described in chapters 5 and 6.

Chapter 4

ENDEAVOUR update: a web resource for gene prioritization in multiple species

4.1 Summary

A first step towards the development of a cross-species gene prioritization method is the extension of the human method to model organisms for which numerous genomic data sources are also available. The choice of model organisms phylogenetically related to human is also motivated by the long term aim of a cross-species prioritization tool. That is, prioritization for one species using data from related organisms. Among the available model organisms, mouse (Mus musculus), rat (Rattus norvegicus), and worm (Caenorhabditis elegans) are chosen because they have been well studied and therefore several genomic data sources are available for these. The extension described in this chapter covers the integration of novel data sources for these model organisms as well as for human, bringing the total to 51 distinct data sources to perform candidate gene prioritization in four organisms. In addition, this chapter also relates the creation of a web based interface that is more intuitive and user friendly. This new interface is simpler since it does not propose all the options that the original interface proposes. This interface is designed to quickly become the default interface while the original interface is reserved for advanced users who need to fine tune their prioritization experiments.

Last point, a pathway based benchmark is performed and confirms the quality of the overall approach (AUC of 88%, 92%, 90% and 86% for human, mouse, rat, and

worm respectively with pathway data sources such as Kegg, Gene Ontology, String and Text excluded). Beside this benchmark, a small scale validation is realized using 32 disease-gene associations that are reported in the literature at least six months after the retrieval of the genomic data (to mimic a real situation in which the association is still unknown). This benchmark revealed that 87,5% (28 genes out of the 32 genes) rank in the top 20%, and that half of these (14 genes) even rank in the top 5%. This confirms one finding of the original publication, *i.e.*, that the performance of the method for real predictive studies is encouraging, although a bit lower than the performance for benchmark studies as expected [167].

Published online 28 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W377–W384***
doi:10.1093/nar/gkn325

ENDEAVOUR update: a web resource for gene prioritization in multiple species

Léon-Charles Tranchevent[1], Roland Barriot[1], Shi Yu[1], Steven Van Vooren[1], Peter Van Loo[1,2,3], Bert Coessens[1], Bart De Moor[1], Stein Aerts[3,4] and Yves Moreau[1,*]

[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, [2]Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven, [3]Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and [4]Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium)

Received February 7, 2008; Revised April 30, 2008; Accepted May 7, 2008

ABSTRACT

ENDEAVOUR (http://www.esat.kuleuven.be/endeavour web; this web site is free and open to all users and there is no login requirement) is a web resource for the prioritization of candidate genes. Using a training set of genes known to be involved in a biological process of interest, our approach consists of (i) inferring several models (based on various genomic data sources), (ii) applying each model to the candidate genes to rank those candidates against the profile of the known genes and (iii) merging the several rankings into a global ranking of the candidate genes. In the present article, we describe the latest developments of ENDEAVOUR. First, we provide a web-based user interface, besides our Java client, to make ENDEAVOUR more universally accessible. Second, we support multiple species: in addition to *Homo sapiens*, we now provide gene prioritization for three major model organisms: *Mus musculus*, *Rattus norvegicus* and *Caenorhabditis elegans*. Third, ENDEAVOUR makes use of additional data sources and is now including numerous databases: ontologies and annotations, protein–protein interactions, *cis*-regulatory information, gene expression data sets, sequence information and text-mining data. We tested the novel version of ENDEAVOUR on 32 recent disease gene associations from the literature. Additionally, we describe a number of recent independent studies that made use of ENDEAVOUR to prioritize candidate genes for obesity and Type II diabetes, cleft lip and cleft palate, and pulmonary fibrosis.

BACKGROUND

With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, converting genomic data into biological knowledge to identify genes involved in a particular process or disease remains a major challenge. Nevertheless, there is much evidence to suggest that functionally related genes often cause similar phenotypes (1–3). To identify which genes are responsible for which phenotype, association studies and linkage analyses are often used, resulting in large lists of candidate genes. In many cases, the list of candidates can be narrowed down to a few dozen. However, it is generally too expensive and time-consuming to perform experimental validation for all these candidates. Therefore, these candidates may be prioritized to first validate the best ones. Given the amount of genomic data publicly available, it is often prohibitive to perform the prioritization manually and consequently, there is a need for computational approaches.

During the past 5 years, the bioinformatics community has developed several strategies to address this question, and several tools are available online (4,5). To our knowledge, all the tools use the concept of similarity. It is based on the assumption that similar phenotypes are caused by genes with similar or related functions (1–3). However, the tools differ by the strategy they adopt in calculating the similarity (either between the candidate genes and the phenotypes or between the candidate genes and the training genes) and by the data sources they use. The most commonly used data sources are text-mining data, gene expression data and sequence information. Additionally, phenotypic data, protein–protein

*To whom correspondence should be addressed. Tel: +32 16 32 17 09; Fax: +32 16 32 19 70; Email: yves.moreau@esat.kuleuven.be

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

interactions, ontologies and *cis*-regulatory information are sometimes included. However, most of the existing approaches mainly focus on the combination of few data sources. For instance, the combining gene expression and protein interaction data method proposed by Ma *et al*. (6) combines expression and interaction data. Several methods only rely on literature and ontologies: BITOLA (7), POCUS (8) Gentrepid (9), G2D (10) and the method defined by Tiffin *et al*. (11). In contrast, systems that use more data sources have recently been designed, such as CAESAR (12), GeneSeeker (13), SUSPECTS (14), TOM (15) and ENDEAVOUR (16). For a more detailed description of the available tools, see the reviews by Oti and Brunner (5) or by Zhu and Zhao (4).

We previously presented the concept of gene prioritization through genomic data fusion and its implementation called ENDEAVOUR (16). This tool requires two inputs: the training genes, already known to be involved in the process under study, and the candidate genes to prioritize. ENDEAVOUR produces one output: the prioritized list of candidate genes, along with the rankings per data source. The algorithm is made up of three stages, called the training, scoring and fusion stages. In the training stage, ENDEAVOUR uses the training genes provided by the user to infer several models, one per data source. For example, with ontology-based data sources, genes are annotated with several terms and reciprocally one term can be associated to several genes. The algorithm selects only the significant terms, the ones that are over-represented in the training sets compared to the complete genome. Hence, the model consists of these significant terms together with their corresponding *P-values* that reflect the significance of the enrichment. In the scoring stage, the model is used to score the candidate genes and rank them according to their score. For ontologies, the algorithm scores each candidate independently by combining the *P-values* of its associated terms that are, at the same time, present in the model. The scores are then used to rank the candidates based on this one data source. In the final stage, the rankings per data source are fused into one global ranking using order statistics. Among the existing methods, the order statistics has the advantage of avoiding penalizing genes that are absent from a given data source. Indeed, the genomic data sources are almost always incomplete. For instance, some genes do not have any ontology annotations, while other genes do not have their corresponding probes spotted on the microarray platform for which data is available. The order statistics allows us to combine the rankings per data source, taking missing values into account. Thus, the use of 'unbiased' data sources (e.g. gene expression data, *cis*-regulatory motifs and protein sequences), together with the use of the order statistics, allows us to obtain results that are not overly biased towards the most studied genes (16). The use of several data sources is indeed an important strength of our approach: combining two data sources, although possibly incomplete, can be more powerful than either individual data source, as shown by our validation experiments (16). The fact that our approach does not rely only on a single data source also reinforces its robustness to noisy data sources like microarray data. More details about the training and scoring methods, the data sources and the order statistics can be found in Supplementary Tables 1 and 2 and in Supplementary Note 1.

In the present article, we describe a novel intuitive web interface in addition to the original Java client. Furthermore, three major model organisms have been added to the application: *M. musculus*, *R. norvegicus* and *C. elegans* (*Danio rerio* and *Drosophila melanogaster* versions will be made available in 2008). Finally, novel data sources have been integrated including numerous protein–protein interaction databases and large species-specific expression data sets, bringing the number of available data sources to 26. Apart from our extensive validation (16), other recent independent publications confirm that ENDEAVOUR is efficient in identifying novel disease genes. Indeed, ENDEAVOUR was recently applied to analyze the adipocyte proteome (17) and to propose novel genes involved in Type II diabetes (18), cleft lip and cleft palate phenotypes (19), and pulmonary fibrosis (20).

OUTLINE OF THE ENDEAVOUR WEB SERVER

ENDEAVOUR was first implemented as a Java client application interacting with a SOAP server and a MySQL database. To make it more universally accessible, we have developed a PHP web-based interface that runs with the most common web browsers, without the need for Java to be installed. It is freely accessible and there is no login requirement.

A four-step wizard guides the user through the preparation of the prioritization (Figure 1). The first step is to choose the organism: human, rat, mouse or worm. The second step is to specify the training set. The user can input a mixture of chromosomal bands, chromosomal intervals, gene symbols, EnsEMBL (21) gene identifiers, KEGG (22) identifiers, Gene Ontology (23) identifiers or OMIM (24) disease names. Each input has to be prefixed according to its type. The rules are explained in the Supplementary material and in the online manual. The genes corresponding to the input are retrieved and loaded into the application. The third step is to select the data sources to be used. The data sources available depend on the organism chosen in the first step. Some of these are species specific (e.g. gene expression data sets) while others are more generic (e.g. Gene Ontology annotations). The last step lets the user specify the candidate genes applying the same rules as in the second step. The user launches the prioritization by using a dedicated button. The computation time is dependent on the number of data sources used, the number of candidates and the load on our servers. The application can handle the prioritization of hundreds of genes (e.g. the average computation time for 400 candidates using 10 data sources is 19.14 s over 100 repeats). Warnings and errors, such as unrecognized gene identifiers, are displayed in the console located in the middle of the main windows. The results are displayed at the bottom of the main page in three panels. The first panel contains the sprint plot, a graphical representation of the rankings with one column per data source plus an additional one for the global ranking. The genes are

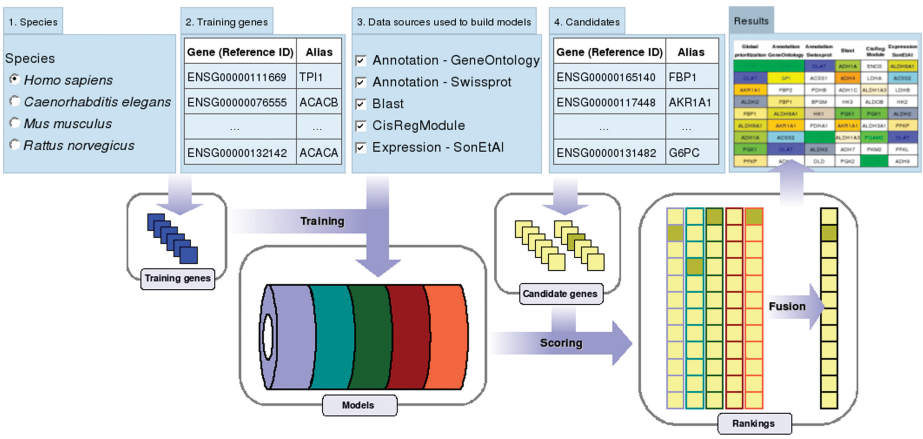

Figure 1. ENDEAVOUR: the algorithm behind the wizard. Once the organism of interest is chosen (Step 1), the user can specify the training genes (Step 2). Step 3 lets the user select the data sources that will be used to build the models. The models summarize the training gene information. The candidate genes specified by the user in Step 4 are then scored against the model. This produces one ranking per data source plus one global ranking obtained by fusion of the rankings per data source. The global ranking together with the rankings per data source are returned to the application and can be viewed in the 'Results' panel.

represented as boxes and the top ranking boxes are coloured for better interpretation of the results. The second panel contains the raw scores and ranks for each gene in each data source. The user can sort the columns according to the global ranking or to any ranking per data source. The third panel allows one to export the results as a TSV spreadsheet or as an XML file. The user can also save the sprint plot using several picture formats (i.e. PNG, JPG and GIF).

NEW MODEL ORGANISMS AND MORE DATA SOURCES

ENDEAVOUR is designed as a generic prioritization tool and is equally useful for the prioritization of candidate disease genes as for candidate members of biological pathways and processes. This is illustrated in our previous publication (16) where we used ENDEAVOUR to identify downstream genes of myeloid differentiation. Since the fundamental study of biological processes is predominantly performed in model organisms, we decided to extend our framework to several model organisms. Currently, gene prioritization can be performed for *M. musculus*, *R. norvegicus* and *C. elegans*, and we are also developing the versions for *D. rerio* and *D. melanogaster*. We have designed the web server so that the organism-specific versions use the same method for each generic data source (e.g. Gene Ontology annotations).

The key strength of ENDEAVOUR resides in the fact that a lot of data sources are available and the user can select the ones that best correspond to the biological question under study. There are 8, 11, 12 and 20 data sources available, respectively, for *R. norvegicus*, *C. elegans*, *M. musculus* and *H. sapiens*, which, in total, result in 26 distinct data sources. They can be classified into six categories: ontologies, interactions, expression, regulatory information, sequence data and text-mining data. Ontologies are structured vocabularies that are used to describe the function of the gene products. Ontologies give more insight on the molecular functions performed [Gene Ontology (23) and SwissProt (25)], on the biological processes involved in [Gene Ontology and KEGG (22)], on the cellular components in which the gene products are active (Gene Ontology) and on the active domains of the proteins [InterPro (26)]. Interaction data come from databases that collect pairs of proteins that interact either physically or genetically. BIND (27) and DIP (28) curate the experimentally determined interactions collected from large-scale interaction and mapping experiments done using yeast two hybrid, mass spectrometry, genetic interactions and phage display. MINT (29) and MIPS (30) mine the literature, either manually or automatically, to find experimentally verified protein interactions. HPRD (31) does the same with an emphasis on domain architecture, post-translational modifications, interaction networks and disease association. IntAct (32) and BioGrid (33) collect physical and genetic interactions by combining analysis of high-throughput experiments and literature curation. STRING (34) and IntNetDb (35) are large databases that contain all kinds of interactions. They rely on a statistical framework to integrate data coming from numerous experiments and databases (including several databases described above), and, additionally, the interactions are transferred across the different organisms, when applicable. Regarding the expression data, the preferred studies are the ones that include a large number of tissues and a large number of genes.

Two sets are available for *H. sapiens* [Su *et al.* (36) and Son *et al.* (37)], three for *M. musculus* [Su *et al.* (36), Hovatta *et al.* (38) and Lindsley *et al.* (39)] and one for *R. norvegicus* and *C. elegans*, respectively from the Walker *et al.* paper (40) and the Baugh *et al.* study (41). Additionally, anatomical expression sequence tags (EST) expression data from EnsEMBL (21) are available for human. Regarding the *cis*-regulatory data, we only have information for *H. sapiens* currently. Using the TOUCAN toolbox (42) and the upstream sequence of the genes, the algorithm looks for putative motifs and modules (combination of five motifs). There are two data sources that are based on sequences: the protein sequence similarities and the disease probabilities. For the latter, Lopez-Bigas *et al.* (43) and Adie *et al.* (44) (ProspectR) used sequence features (e.g. length of the sequence, length of the UTRs, number of introns, length of the introns) and a statistical framework to discriminate the human disease causing genes from the rest of the genome. Next, they associated to every gene a probability of being a disease causing gene, *a priori*. As for sequence similarity, an all-against-all similarity search is performed for all organisms using the NCBI BLAST (45). The data source based on literature mining relies on the TxtGate framework (46). The strategy is to screen the abstracts from PubMed (47) with a manually curated vocabulary based on Gene Ontology. Similarly to the ontologies described above, it provides more information on the molecular functions and biological processes of the genes. It is important to notice that, except for the regulatory information category, each organism is provided with at least one data source per category.

As an alternative to the novel web-based application, one can use the original Java Web Start client, which is also extended to include the other model organisms. This application includes a few additional features, such as a full description of the models created, a full genome screening service in which the whole genome of the given organism can be prioritized and the possibility for users to make use of their own microarray data sets. A SOAP service is also available to allow integration in workflows [e.g. when using Taverna (48) or Kepler (49)].

SOFTWARE DOCUMENTATION

ENDEAVOUR comes with an online manual. A subsection describes the concept of gene prioritization through genomic data fusion. Another subsection contains the answers to frequently asked questions and gives more details on how to perform a prioritization and how to interpret the results. Finally, a step-by-step example is given together with the corresponding screenshots.

The application is provided with three use cases taken from the literature. The user can run the examples by clicking on the corresponding buttons situated above the wizard that cause the training genes, the data sources and the candidate genes to be loaded automatically into the application. Then, the user can quickly go through the four steps and launch the prioritization process. The three use cases can be used as a first step to understand the mechanisms of ENDEAVOUR. The first example is derived from our previous publication in which we studied the DiGeorge syndrome (16). This example shows why *YPEL1* was first selected for wet lab experiments that eventually confirmed the phenotypic association in zebrafish. The second example is taken from the Elbers *et al.* (18) review on obesity and Type II diabetes. They have prioritized five susceptibility loci to reveal a molecular link between the two disorders. ENDEAVOUR uncovered the susceptibility loci located on chromosome 11 for this example. It contains *KCNJ5*, a homolog of *KCNJ11* that is known to contribute to the risk of Type II diabetes. We have built the last example after Ebermann *et al.* (50) published their discovery of a novel Usher gene, *DFNB31*, that encodes the whirlin protein. By using data six months prior to the publication, we made sure that the association was not yet present in the databases. Among the 32 candidates of the chromosomal band 9q32, *DFNB31* ranked first, showing that, retrospectively, it was indeed a good candidate.

VALIDATION

Similarly to our previous work (16), we statistically validate the approach with a standard leave-one-out cross-validation using known gene sets. We produced the corresponding receiver operating characteristic (ROC) curves and measured the performance by calculating the area under the curve (AUC) (Figure 2). Here, we focused on the pathway gene prioritization for the newly added species by applying this scheme to three signalling pathways taken from the Gene Ontology database (23). These pathways are common to the four organisms and involve, respectively, 193, 170, 126 and 44 genes for *H. sapiens*, *M. musculus*, *R. norvegicus* and *C. elegans*. We performed both a fair validation and a complete validation. For the fair validation, we excluded the data sources that might contain explicitly the gene-pathway association (i.e. Gene Ontology, Kegg, String and Text) while all data sources were used for the complete validation. The first observation is that the performance of the four control validations stays close to the theoretical expectation of 50% (respectively, 48, 39, 45 and 51%). This means that when using randomly generated gene sets for training, we obtain random results. In contrast, the performance of biologically meaningful sets is much higher (respectively, 88, 92, 90 and 86% for the fair validation and 99, 99, 99 and 98% for the complete validation). An analysis per data source of the fair validation reveals that the global performance (e.g. 88% for human) is always higher than the best performing data source performance (e.g. 78% for human InterPro). It shows that our data fusion approach is scientifically sound and that it is crucial to make use of complementary data sources. Altogether, this indicates that our approach based on the assumption that functionally related genes often cause similar phenotypes can be applied successfully.

A difficulty of validating gene prioritization methods is the fact that known data are used for the ranking. In other words, for every disease or pathway gene, the link between the disease and the gene is described in the literature and

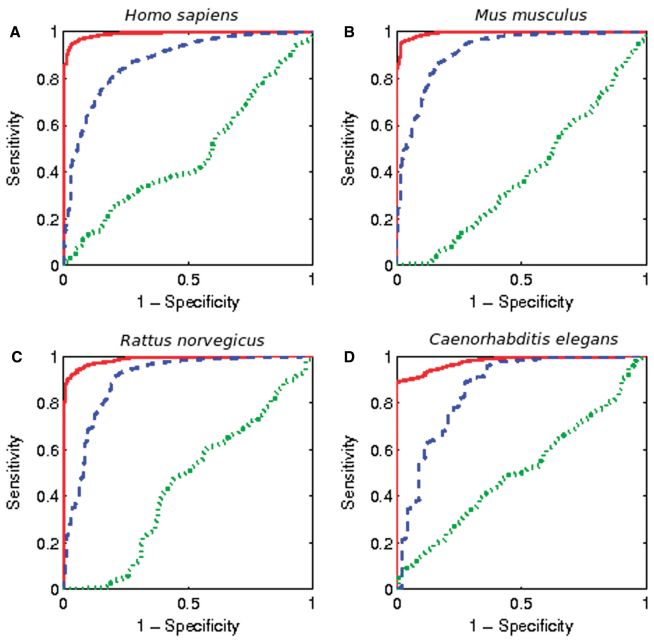

Figure 2. Results of the leave-one-out cross-validation. For each organism, the leave-one-out cross-validation was performed on three pathways sets from Gene Ontology (23), and, as a control, on five sets of 20 randomly selected genes. The ROC curves of the random (dotted green) and pathway validation (solid red and dashed blue) are plotted for (**a**) *H. sapiens*, (**b**) *M. musculus*, (**c**) *R. norvegicus* and (d) *C. elegans*. Notice that for the fair validation (dashed blue), Gene Ontology, KEGG, Text and String were excluded while all data sources were used for the complete validation (solid red). The AUC of the control validations are respectively 48, 39, 45 and 51% indicating a random performance. On the opposite, the AUC of the pathway validations are respectively 88, 92, 90 and 86% for the fair validation and 99, 99, 99 and 98% for the complete validation showing the validity of our approach.

sometimes evidence is also present in the ontologies or in the interaction information. Therefore, we excluded in the above analysis the data sources that contain explicit information about the similarity of the true positive to the training set. To assess the full performance of ENDEAVOUR to solve real biological cases, using all data sources, we therefore focused on genetic disorders for which associations were reported very recently in the literature, so that the explicit information is not yet present in our data. Particularly, we used gene–disease associations that were reported in *Nature Genetics* after 1 January 2008 (Table 1), 32 in total. For each disorder, we built a training set containing all the genes already known to play a role in that disorder according to the OMIM and Gene Ontology databases (both downloaded in August 2007). As candidate genes to be ranked we used the true positive gene together with 99 genes that flank the true positive in the genome. These regions were then prioritized with ENDEAVOUR using all data sources and their specific training sets. The results are presented in Table 1. Interestingly, *BANK1*, *CTRC* and *SORT1* rank first out of their region and *GDF5*, *RGS1* and *SH2B3* rank second. All genes but four are within the top 20% and half of them are within the top 9%.

Others have used our gene prioritization tool as well. Elbers *et al.* (18) have used ENDEAVOUR in combination with other prioritization tools to define the best strategy to search for common obesity and Type II diabetes genes. They suggest a list of genes indicated as potential candidates by at least two of the six tools. Tzouvelekis *et al.* (20) have used ENDEAVOUR to prioritize a list of genes differentially expressed in idiopathic pulmonary fibrosis. They consistently find that among the top candidates, five and seven genes are targets of, respectively, tumor necrosis factor (TNF) and transforming growth factor (TGF). Osoegawa *et al.* (19) applied ENDEAVOUR to propose novel genes associated with cleft lip and cleft palate phenotypes. They analysed 83 syndromic cases and 104 non-syndromic cases and concluded that estrogen receptor 1 (ESR1) and fibroblast growth factor receptor 2 (FGFR2) were the most likely candidates, respectively, from region 6q25.1-25.2 and region 10q26.11-26.13. Using mass spectrometry and bioinformatics, Adachi *et al.* (17) explored the proteome of the adipocyte, a central player in energy metabolism.

Table 1. Results of the thirty two genetic disorder prioritizations

Gene	Disorder	Reference	Endeavour rank
BANK1	Systemic lupus erythematosus	Kozyrev *et al.* (51)	1
ITGAM	Systemic lupus erythematosus	Nath *et al.* (52)	3
TNFSF4	Systemic lupus erythematosus	Graham *et al.* (53)	16
DPP6	Amyotropic lateral sclerosis	van Es *et al.* (54)	15
CTRC	Chronic pancreatitis	Rosendahl *et al.* (55)	1
ATP6V0A2	Impaired glycosylation	Kornak *et al.* (56)	5
ATP6V0A2	Cutis laxa	Kornak *et al.* (56)	5
GALNT2[a]	LDL/HDL cholesterol	Willer *et al.* (57), Kathiresan *et al.* (58)	13
SORT1[a]	LDL/HDL cholesterol	Willer *et al.* (57), Kathiresan *et al.* (58)	1
MLXIPL[a]	LDL/HDL cholesterol	Willer *et al.* (57), Kathiresan *et al.* (58), Kooner *et al.* (59),	12
GDF5[a]	Human height	Sanna *et al.* (60)	2
C20orf44[a]	Human height	Sanna *et al.* (60)	41
MSMB[a]	Prostate cancer	Eeles *et al.* (61), Thomas *et al.* (62)	18
JAZF1[a]	Prostate cancer	Thomas *et al.* (62)	14
CTBP2[a]	Prostate cancer	Thomas *et al.* (62)	4
LMTK2[a]	Prostate cancer	Eeles *et al.* (61)	4
KLK3[a]	Prostate cancer	Eeles *et al.* (61)	9
CPNE3[a]	Prostate cancer	Thomas *et al.* (62)	42
IL16[a]	Prostate cancer	Thomas *et al.* (62)	9
CDH23[a]	Prostate cancer	Thomas *et al.* (62)	40
EHBP1[a]	Prostate cancer	Gudmundsson *et al.* (63)	19
CCR3[a]	Celiac disease	Hunt *et al.* (64)	12
RGS1[a]	Celiac disease	Hunt *et al.* (64)	2
LPP[a]	Celiac disease	Hunt *et al.* (64)	30
TAGAP[a]	Celiac disease	Hunt *et al.* (64)	3
SH2B3[a]	Celiac disease	Hunt *et al.* (64)	2
IL12A[a]	Celiac disease	Hunt *et al.* (64)	18
SCHIP1[a]	Celiac disease	Hunt *et al.* (64)	20
IL18R1[a]	Celiac disease	Hunt *et al.* (64)	3
IL18RAP[a]	Celiac disease	Hunt *et al.* (64)	4
IL2[a]	Celiac disease	Hunt *et al.* (64)	10
IL21[a]	Celiac disease	Hunt *et al.* (64)	14
		Mean (all genes)	12.25
		Mean (GWAS excluded)	6.57

[a]Associations reported with GWAS (Genome Wide SNPs Associations Studies).
The gene-disease associations were reported in Nature Genetics after 1 January 2008 to exclude the presence of explicit evidence in our data sources. The training sets were built with OMIM and Gene Ontology; and the candidate regions contain the novel gene and its 99 nearest neighbours. The 20 human data sources were used to perform the prioritizations. The results show that ENDEAVOUR ranked all the novel genes but four within the top 20%, and half of them within the top 9%.

Using ENDEAVOUR, they were able to associate a number of factors with vesicle transport in response to insulin stimulation, which is a key function of adipocytes.

CONCLUSION

ENDEAVOUR is a web server that allows users to prioritize candidate genes with respect to their biological processes or diseases of interest. It is provided with an intuitive four-step wizard and an online manual. It is available for four organisms (*H. sapiens*, *M. musculus*, *R. norvegicus* and *C. elegans*). ENDEAVOUR relies on the similarity between the candidates and the models built with the training genes. The approach has been validated experimentally (16), by extensive leave-one-out cross-validations, and by analysis of recently reported cases from the literature. Additionally, several independent laboratories have used ENDEAVOUR to propose novel disease genes [Elbers *et al.* (18) and Osoegawa *et al.* (19)] or to optimize the analysis of medium-throughput experiments [Tzouvelekis *et al.* (20) and Adachi *et al.* (17)]. Importantly, the cross-validation revealed the added value of combining several complementary data sources. With 26 distinct data sources (51 in total) covering most aspects of the knowledge available on genes and gene products (functional annotations, protein interactions, expression profiles, regulatory information, sequence-based data and literature mining), ENDEAVOUR exploits the most comprehensive collection of publicly available knowledge.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This research was supported by the Research Council KUL (GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, several PhD/postdoc & fellow grants), FWO [PhD/postdoc grants, projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM)], IWT (PhD Grants, GBOU-McKnow-E (Knowledge management

algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM Endometriosis), the Belgian Federal Science Policy Office [IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011), and the EU-RTD (ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap)]. The authors thank Sonia Leach for critical comments and helpful suggestions on the article. P.V.L. and S.A. are, respectively, supported by a PhD and a postdoctoral research fellowship of the Research Foundation—Flanders (FWO).

Conflict of interest statement. None declared.

REFERENCES

1. Smith,N.G. and Eyre-Walker,A. (2003) Human disease genes: patterns and predictions. *Gene*, **318**, 169–175.
2. Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabási,A.L. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.
3. Jimenez-Sanchez,G., Childs,B. and Valle,D. (2001) Human disease genes. *Nature*, **409**, 853–855.
4. Zhu,M. and Zhao,S. (2007) Candidate gene identification approach: progress and challenges. *Int. J. Biol. Sci.*, **3**, 420–427.
5. Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.
6. Ma,X., Lee,H., Wang,L. and Sun,F. (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, **23**, 215–221.
7. Hristovski,D., Peterlin,B., Mitchell,J.A. and Humphrey,S.M. (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.*, **74**, 289–298.
8. Turner,F.S., Clutterbuck,D.R. and Semple,C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.
9. George,R.A., Liu,J.Y., Feng,L.L., Bryson-Richardson,R.J., Fatkin,D. and Wouters,M.A. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.
10. Perez-Iratxeta,C., Wjst,M., Bork,P. and Andrade,M.A. (2005) G2D: a tool for mining genes associated with disease. *BMC Genet.*, **6**, 45.
11. Tiffin,N., Kelso,J.F., Powell,A.R., Pan,H., Bajic,V.B. and Hide,W.A. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, **33**, 1544–1552.
12. Gaulton,K.J., Mohlke,K.L. and Vision,T.J. (2007) A computational system to select candidate genes for complex human traits. *Bioinformatics*, **23**, 1132–1140.
13. van Driel,M.A., Cuelenaere,K., Kemmeren,P.P., Leunissen,J.A., Brunner,H.G. and Vriend,G. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.*, **33**, W758–W761.
14. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.
15. Rossi,S., Masotti,D., Nardini,C., Bonora,E., Romeo,G., Macii,E., Benini,L. and Volinia,S. (2006) TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res.*, **34**, W285–W292.
16. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.-C., De Moor,B., Marynen,P., Hassan,B. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.
17. Adachi,J., Kumar,C., Zhang,Y. and Mann,M. (2007) In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol. Cell. Proteomics*, **6**, 1257–1273.
18. Elbers,C., Onland-Moret,C., Franke,L., Niehoff,A., van der Schouw,Y. and Wijmenga,C. (2007) A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol. Metab.*, **18**, 19–26.
19. Osoegawa,K., Vessere,G., Utami,K., Mansilla,M., Johnson,M., Riley,B., L'Heureux,J., Pfundt,R., Staaf,J., van der Vliet,W. *et al.* (2008) Identification of novel candidate genes associated with cleft lip and palate using array comparative genomic hybridisation. *J. Med. Genet.*, **45**, 81–86.
20. Tzouvelekis,A., Harokopos,V., Paparountas,T., Oikonomou,N., Chatziioannou,A., Vilaras,G., Tsiambas,E., Karameris,A., Bouros,D. and Aidinis,V. (2007) Comparative expression profiling in pulmonary fibrosis suggests a role of hypoxia-inducible factor-1alpha in disease pathogenesis. *Am. J. Respir. Crit. Care Med.*, **176**, 1108–1119.
21. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
22. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
23. The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
24. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
25. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
26. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
27. Bader,G., Donaldson,I., Wolting,C., Ouellette,F., Pawson,T. and Hogue,C. (2001) BIND-The biomolecular interaction network database. *Nucleic Acids Res.*, **29**, 242–245.
28. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
29. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the molecular interaction database. *Nucleic Acids Res.*, **35**, D572–D574.
30. Mewes,H.W., Frishman,D., Mayer,K.F.X., Munsterkotter,M., Noubibou,O., Pagel,P., Rattei,T., Oesterheld,M., Ruepp,A. and Stumpflen,V. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.
31. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
32. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct-open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
33. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
34. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7-recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
35. Xia,K., Dong,D. and Han,J.D. (2006) IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, **7**, 508.
36. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
37. Son,C.G., Bilke,S., Davis,S., Greer,B.T., Wei,J.S., Whiteford,C.C., Chen,Q.R., Cenacchi,N. and Khan,J. (2005) Database of mRNA

gene expression profiles of multiple human organs. *Genome Res.*, **15**, 443–450.
38. Hovatta,I., Tennant,R.S., Helton,R., Marr,R.A., Singer,O., Redwine,J.M., Ellison,J.A., Schadt,E.E., Verma,I.M., Lockhart,D.J. *et al.* (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature*, **438**, 662–666.
39. Lindsley,R.C., Gill,J.G., Kyba,M., Murphy,T.L. and Murphy,K.M. (2006) Canonical Wnt signaling is required for development of embryonic stem cell-derived mesoderm. *Development*, **133**, 3787–3796.
40. Walker,J.R., Su,A.I., Self,D.W., Hogenesch,J.B., Lapp,H., Maier,R., Hoyer,D. and Bilbe,G. (2004) Applications of a rat multiple tissue gene expression data set. *Genome Res.*, **14**, 742–749.
41. Baugh,L.R., Hill,A.A., Claggett,J.M., Hill-Harfe,K., Wen,J.C., Slonim,D.K., Brown,E.L. and Hunter,C.P. (2005) The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. *Development*, **132**, 1843–1854.
42. Aerts,S., Van Loo,P., Thijs,G., Mayer,H., de Martin,R., Moreau,Y. and De Moor,B. (2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.*, **33**, W393–W396.
43. Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, **32**, 3108–3114.
44. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.
45. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.
46. Glenisson,P., Coessens,B., Van Vooren,S., Mathys,J., Moreau,Y. and De Moor,B. (2004) TXTGate: profiling gene groups with text-based information. *Genome Biol.*, **5**, R43.
47. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
48. Oinn,T., Addis,M., Ferris,J., Marvin,D., Senger,M., Greenwood,M., Carver,T., Glover,K., Pocock,M.R., Wipat,A. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.
49. Altintas,I., Berkley,C., Jaeger,E., Jones,M., Ludäscher,B. and Mock,S. (2004) *16th International Conference on Scientific and Statistical Database Management* Santorini Island, Greece.
50. Ebermann,I., Scholl,H.P., Charbel Issa,P., Becirovic,E., Lamprecht,J., Jurklies,B., Millan,J.M., Aller,E., Mitter,D. and Bolz,H. (2007) A novel gene for Usher syndrome type 2: mutations in the long isoform of whirlin are associated with retinitis pigmentosa and sensorineural hearing loss. *Hum. Genet.*, **121**, 203–211.
51. Kozyrev,S.V., Abelson,A.K., Wojcik,J., Zaghlool,A., Linga Reddy,M.V., Sanchez,E., Gunnarsson,I., Svenungsson,E., Sturfelt,G., Jönsen,A. *et al.* (2008) Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat. Genet.*, **40**, 211–216.
52. Nath,S.K., Han,S., Kim-Howard,X., Kelly,J.A., Viswanathan,P., Gilkeson,G.S., Chen,W., Zhu,C., McEver,R.P., Kimberly,R.P. *et al.* (2008) A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. *Nat. Genet.*, **40**, 152–154.
53. Graham,D.S., Graham,R.R., Manku,H., Wong,A.K., Whittaker,J.C., Gaffney,P.M., Moser,K.L., Rioux,J.D., Altshuler,D., Behrens,T.W. *et al.* (2008) Polymorphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus. *Nat. Genet.*, **40**, 83–89.
54. van Es,M.A., van Vught,P.W., Blauw,H.M., Franke,L., Saris,C.G., Van den Bosch,L., de Jong,S.W., de Jong,V., Baas,F., van't Slot,R. *et al.* (2008) Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.*, **40**, 29–31.
55. Rosendahl,J., Witt,H., Szmola,R., Bhatia,E., Ozsvári,B., Landt,O., Schulz,H.U., Gress,T.M., Pfützer,R., Löhr,M. *et al.* (2008) Chymotrypsin C (CTRC) variants that diminish activity or secretion are associated with chronic pancreatitis. *Nat. Genet.*, **40**, 78–82.
56. Kornak,U., Reynders,E., Dimopoulou,A., van Reeuwijk,J., Fischer,B., Rajab,A., Budde,B., Nürnberg,P., Foulquier,F., ARCL Debré-type Study Group. *et al.* (2008) Impaired glycosylation and cutis laxa caused by mutations in the vesicular H+-ATPase subunit ATP6V0A2. *Nat. Genet.*, **40**, 32–34.
57. Willer,C.J., Sanna,S., Jackson,A.U., Scuteri,A., Bonnycastle,L.L., Clarke,R., Heath,S.C., Timpson,N.J., Najjar,S.S., Stringham,H.M. *et al.* (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.*, **40**, 161–169.
58. Kathiresan,S., Melander,O., Guiducci,C., Surti,A., Burtt,N.P., Rieder,M.J., Cooper,G.M., Roos,C., Voight,B.F., Havulinna,A.S. *et al.* (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.*, **40**, 189–197.
59. Kooner,J.S., Chambers,J.C., Aguilar-Salinas,C.A., Hinds,D.A., Hyde,C.L., Warnes,G.R., Gómez,P.J., Frazer,K.A., Elliott,P., Scott,J. *et al.* (2008) Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat. Genet.*, **40**, 149–151.
60. Sanna,S., Jackson,A.U., Nagaraja,R., Willer,C.J., Chen,W.-M., Bonnycastle,L.L., Shen,H., Timpson,N., Lettre,G., Usala,G. *et al.* (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.*, **40**, 198–203.
61. Eeles,R.A., Kote-Jarai,Z., Giles,G.G., Olama,A.A.A., Guy,M., Jugurnauth,S.K., Mulholland,S., Leongamornlert,D.A., Edwards,S.M., Morrison,J. *et al.* (2008) Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.*, **40**, 316–321.
62. Thomas,G., Jacobs,K.B., Yeager,M., Kraft,P., Wacholder,S., Orr,N., Yu,K., Chatterjee,N., Welch,R., Hutchinson,A. *et al.* (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, **40**, 310–315.
63. Gudmundsson,J., Sulem,P., Rafnar,T., Bergthorsson,J.T., Manolescu,A., Gudbjartsson,D., Agnarsson,B.A., Sigurdsson,A., Benediktsdottir,K.R., Blondal,T. *et al.* (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.*, **40**, 281–283.
64. Hunt,K.A., Zhernakova,A., Turner,G., Heap,G.A.R., Franke,L., Bruinenberg,M., Romanos,J., Dinesen,L.C., Ryan,A.W., Panesar,D. *et al.* (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.*, **40**, 395–402.

4.2 Contribution of the PhD candidate

The PhD candidate has gathered the genomic data for the extra three species. He has also updated the already existing data sources accordingly. He has contributed to the development of the web interface through the co-development of an XML based client-server communication module. In addition, the PhD candidate has performed the benchmark analysis based on OMIM, Gene Ontology and recently reported disease-gene associations. He has written the paper.

4.3 Discussion

Besides mouse, rat and worm, a fruit fly (Drosophila melanogaster) version was further implemented and validated. This work is described in chapter 8. In addition, the tool has recently been extended to zebrafish as well (Danio rerio). Altogether, this corresponds to 26 additional data sources (14 for fruit fly and 12 for zebrafish).

Tools have been integrated to measure how many people are consulting our web interface, this was done using the Google Analytics suite. A summary is described in figure 4.1. There have been a total of 8505 visits between the official launch of the website in April 2008 and April 2010. That represents an average of 82 visits per week. Interested researchers are mainly coming from the United States, China, and Belgium, although other european countries active in bioinformatics have also shown some interest (*e.g.*, France, United Kingdom, Germany). The figure 4.1 also presents the main events that have contributed to increase the number of visits such as the original publication, conferences, or courses in which the tool was demonstrated. Although very modest, these numbers are encouraging and have resulted into several independent publications. The table 4.1 presents the 4 examples already discussed in the current publication as well as 16 additional examples found in the literature since then. Most of these publications represent independent use of our software Endeavour and some have led to some breakthrough in human genetic. They are all described in details below.

4.3.1 External validations

Selection of genes and single nucleotide polymorphisms for fine mapping starting from a broad linkage region Windelinckx *et al.* describe an empirical two-step fine mapping approach, in which candidate genes are prioritized using Endeavour, and the top genes are chosen for further SNP selection with a linkage disequilibrium based method (Tagger) [258]. The authors apply this approach on two previously identified linkage regions for muscle strength. This results in the selection of 331 polymorphisms located in 112 different candidate genes out of an

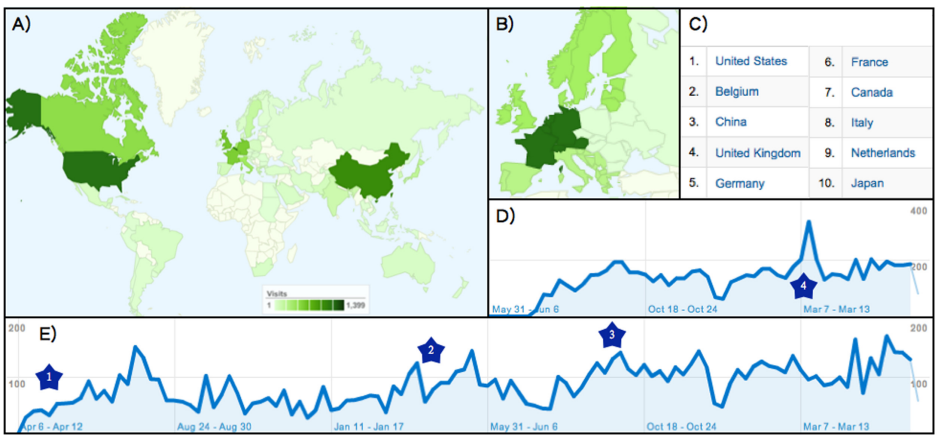

Figure 4.1: Traffic and statistics for the Endeavour website. The statistics are obtained for the prioritization interface (A, B, and E), the main page (D), and the entire website (C) through the use of Google Analytics. (A) World map of the traffic between April 2008 and April 2010 for the prioritization interface, with the United states being the best contributor with 1400 visits. (B) A detailed view of Europe for the same page and covering the same period, showing the best europeans contributors. (C) The 10 most contributing countries, worldwide, results are merged for the entire website. (D) The traffic on the main page between April 2009 and April 2010. (E) The traffic on the prioritization interface between April 2008 and April 2010. The four stars represent events that have contributed to an increase of the traffic. (1) Publication of the web based interface on May 2008. (2) Presentation at the Genomic Disorders conference on Mars 2009. (3) Course in master of Biomedicine at the K.U.L. on September 2009. (4) Entry on the BioMed Central blog (`http://blogs.openaccesscentral.com/blogs/bmcblog/entry/chdwiki_genomemedicine`) mentioning the CHDWiki publication. CHDWiki includes a gene prioritization module (see also chapter 8).

initial set of 23,300 SNPs. Notice that gene prioritization is performed five times using five different training sets all related to muscle strength, and that results are further integrated in order to obtain a global ranking. After prioritization, a total of 129 genes are retained from the 597 genes that lay in the two linkage regions. This is further narrowed down to 331 SNPs using a SNP selection strategy.

A bivariate whole genome linkage study identified genomic regions influencing both BMD and bone structure Liu *et al.* use six gene prioritization tools in conjunction to propose new candidate for areal BMD (aBMD) and areal bone size (ABS) that are both risk factors for osteoporosis [139]. They screen seven linkage

regions associated with both aBMD and ABS and identify 34 promising candidate genes that are reported by at least three of the six prioritization tools.

A new dominantly inherited pure cerebellar ataxia, SCA 30 Storey *et al.* report the analysis of a family manifesting a cerebellar ataxia and identify a locus on 4q34.3-q35.1 associated to the disease [224]. A gene prioritization of the region that encompasses 19 genes reveals that the gene ODZ3 seems an interesting candidate, which is confirmed through further expression analysis that reveals adult and fetal brain expression (very high expression in amygdala and caudate nucleus).

Identification of LTBP2 on chromosome 14q as a novel candidate gene for bone mineral density variation and fracture risk association Cheung *et al.* describe a thorough analysis of a linkage pick at 14q in relation to Bone Mineral Density (BMD) and fracture risk association [53]. More than 300 genes are located in the linkage region, only five are retained after prioritization and are further screened for SNPs. Among the 55 SNPs, 16 show significant associations, 6 of which are in the LTBP2 gene. In addition, LTBP2 expression is observed in preosteoblast cells during osteoblastic proliferation and subsequent differentiation suggesting that LTBP2 might be clinically important in fracture management and therefore in osteoporosis.

The complexity of genotypic alterations underlying HER2-positive breast cancer: an explanation for its clinical heterogeneity Vanden Bempt *et al.* investigate the impact of HER2 gene amplification on the biology of breast tumors [30]. They first define regions of interest by comparative expressed sequence hybridization analysis, they find four regions on chromosomal regions 17q12, 3q24-q26.3, 14q24-31 and 20q12-q13.1 encompassing more than 500 genes each. Gene prioritization is performed to narrow this huge list down to eight genes that are further investigated by quantitative real-time polymerase chain reaction (qRT-PCR). Two out of these eight genes show significant expression in HER2 amplified breast carcinomas (HER2 and CRK7) and a trend can be observed for a third gene (MMP9), although it does not reach significance, suggesting interesting candidate genes for HER2 amplified breast tumors.

Recurrent copy number changes in mentally retarded children harbour genes involved in cellular localization and the glutamate receptor complex Poot *et al.* identify recurrent Copy Number Changes (CNCs) in mentally retarded children and propose novel candidate from these CNCs [191]. Starting from 278 patients and 48 controls and using the array CGH technology, 27 CNCs are identified. Gene set enrichment analysis, gene prioritization and network based analysis are then realized using Gather, Endeavour and Prioritizer, and String. The results show

that genes involved in potassium ion transport and establishment of localization represent promising candidate genes.

Role of genetic variation in insulin-like growth factor 1 receptor on insulin resistance and arterial hypertension Sookoian *et al.* perform a two-stage study to explore the role of gene variants in the risk of insulin resistance and arterial hypertension (in relation with type 2 diabetes) [221]. A whole genome prioritization identified 10 highly significant genes, among which only one (IGF1R) demonstrates significant association in various Genome Wide Association studies (GWAS). Further candidate gene association studies show non significant association between metabolic syndromes and IGF1R. The authors suggest that it might be due to the low Minor Allele Frequency (MAF) of the patient cohort.

Gene prioritization based on biological plausibility over genome wide association studies renders new loci associated with type 2 diabetes Sookoian *et al.* combine two GWAS for type 2 diabetes with prioritization by Endeavour [222]. Starting from the whole genome, 241 genes are prioritized with significant p-values, corresponding to a total of 1096 SNPs. The analysis of the GWAS results show that 6 SNPs are associated with type 2 diabetes, corresponding to five genes: TACR3, ALK, CACNA1D, FOXO1A, and AKT3. The authors report that, in animal models, the haploinsufficiency of the FOXO1A restores insulin sensitivity and rescues the diabetic phenotype in insulin-resistant mice and that, conversely, a gain-of-function FOXO1A mutation results in diabetes. Furthermore, FOXO1A is regarded as a potential therapeutic target for improving insulin resistance. The other candidates are involved in energy balance (ALK), regulation of insulin secretion (CACNA1D), regulation of cell signaling in response to insulin and glucose uptake (AKT3).

The RNA-binding protein Elavl1/HuR is essential for placental branching morphogenesis and embryonic development Katsanou *et al.* are coupling gene expression measurements, *in silico* data analysis, and HuR-RNA association assays to identify transcription and growth factor mRNAs controlled by HuR in rodent embryonic development[121]. First a set of 395 differentially expressed mouse genes is collected from HuR^{-} and HuR^{+} embryos. These genes are further prioritized with Endeavour using three different training sets that describe respectively placental insufficiencies, limb deformities, and asplenia. Only the 23 genes that are significantly prioritized by the three training sets are retained for the qRT-PCR experiments. The results show that 19 out of the 23 genes are indeed differentially expressed. As a last validation, Immuno-Precipitation (IP) assays are run, 6 of the 19 genes are shown to associate distinctly with HuR, among which five have not been reported in the literature before, enriching therefore the role of HuR in mouse embryonic development.

Phenomic determinants of genomic variation in autism spectrum disorders Qiao *et al.* aim at finding Copy Number Variants (CNVs) associated to autism spectrum disorders through the use of the array CGH technology[195]. They identify 12 positive regions, among which 6 contain 3 genes or less. The remaining 6 regions that encompass 187 genes altogether were then prioritized individually and the first 2 genes from each regions were retained. These genes were mostly involved in either mammalian nervous system development and/or neuronal excitability. For example, WNK3, and NALCN have been reported to be associated with neurodevelopmental disorders including NIPA1 (involved in hereditary spastic paraplegia), and CTNND2 (cri-du-chat syndrome). Some have been reported in ASD related studies, such as PHF8, WNK3 SEMA5A, GTF2I, STX1A, NIPA1 and UBE3A.

Identification of neuroglycan C and interacting partners as potential susceptibility genes for schizophrenia in a Southern Chinese population So *et al.* analyze a previously reported linkage region for schizophrenia on chromosome 3p [220]. Endeavour is used to prioritize the 129 genes from that region, and results are combined with the ones of the association screen. At the end, NGC appears as the best candidate (1st rank), its brain specific expression pattern and its involvement in neurodevelopment also support this hypothesis.

Genetic modification of the inner ear lateral semicircular canal phenotype of the Bmp4 haplo-insufficient mouse Vervoort *et al.* use mice to study ear development, more precisely the inner ear lateral semicircular canal phenotype [244]. A genome scan identifies two modifier loci on chromosome 4 and 14. Only the locus on chromosome 4 undergoes prioritization, six genes are selected for further experiments. Only one of the associated SNPs is significant and corresponds to the gene Prdm16. The authors precise that functional assays are still needed to confirm its role in inner ear development.

Narrowing the critical deletion region for autism spectrum disorders on 16p11.2 Crepel *et al.* report a small 118 kb deletion within the recurrent 16p11.2 copy number variant (CNV), segregating with ASD or ASD traits in a three-generation family [57]. The prioritization of the complete recurrent CNV revealed that 2 of the top 5 genes, SEZ6L2 and MPV, are located within the small 118kb deletion. They appear as interesting candidate genes with supporting evidence from human and model organisms.

NEK1 mutations cause short-rib polydactyly syndrome type majewski Thiel *et al.* used homozygosity mapping in two families with autosomal-recessive short-rib polydactyly syndrome Majewski type to identify mutations [230]. They

Publication	Disease / process	Main finding
*Elbers *et al.*	Type 2 diabetes	27 genes[c]
*Tzouvelekis *et al.*	Pulmonary fibrosis	HIF1A
*Osoegawa *et al.*	Cleft lip and palate	ESR1 and FGFR2
*Adachi *et al.*	Energy metabolism	41 genes[c]
Windelinckx *et al.*	Muscle strength	112 genes
Liu *et al.*	Bone mineral density	34 genes[c]
Storey *et al.*	Cerebellar ataxia	ODZ3
Cheung *et al.*	Bone mineral density	TLBP2
Vanden Bempt *et al.*	Breast cancer	HER2, CRK7, and MMP9[a]
Poot *et al.*	Mental retardation	34 genes[c]
Sookoian *et al.*	Insulin resistance	IGF1R[a]
Sookoian *et al.*	Type 2 diabetes	TACR3, ALK, CACNA1D, FOXO1A, and AKT3
Katsanou *et al.*	Embryonic development	Ets2[b], Fgf10[b], Hoxd13[b], Tbx4[b], Foxc1[b], and Hoxb9[b]
Qiao *et al.*	Autism	WNK3, NALCN, CTNND2, PHF8, WNK3, SEMA5A, GTF2I, STX1A, NIPA1 and UBE3A
So *et al.*	Schizophrenia	NGC
Vervoort *et al.*	Ear development	Prdm16[ab]
Crepel *et al.*	autism	SEZ6L2, MPV[a]
Thiel *et al.*	Short rib polydactyly syndrome Majewski	NEK1
Dupé *et al.*	Holoprosencephaly	DLL1
Tanaka *et al.*	Diabetes (MODY)	GCKR[a]

Table 4.1: External validation of Endeavour. List of the 20 external publications that use Endeavour to prioritize candidate genes for various genetic disorders and developmental processes. (*) Publication is reported in Tranchevent *et al.* (2008). ([a]) The reported association is not significant but authors show evidence that gene is still a good candidate. ([b]) Experiments are conducted in mouse and therefore the reported gene is a mouse gene. ([c]) The complete gene lists can be found in appendix B.

identified one linked interval (LOD score 2.95) representing 17.36 Mb/18.65 cM on chromosome 4, encompassing 38 genes. To prioritize these genes under the hypothesis of a defect in cilia function, they used known genes of the cilia proteome database and compared them with the genes from the candidate interval using Endeavour. The NEK1 gene ranks second and mutations in that gene were found in the two families.

NOTCH, a new signaling pathway implicated in holoprosencephaly Dupé *et al.* have analyzed 4 holoprosencephaly (HPE) patients through array CGH, and have defined a redundant 6qter deletion [69]. Through prioritization, the DLL1 gene was identified as the best candidate gene from the region. Expression analysis and mutation screens indeed showed that DLL1 has is involved in early patterning of the forebrain and suggested NOTCH as a new signaling pathway involved in HPE.

GCKR mutations in Japanese families with clustered type 2 diabetes Tanaka *et al.* recruited Japanese families with a 3-generation history of diabetes [227]. Genome-wide linkage analysis was performed assuming an autosomal dominant model. Genes in the linkage region were computationally prioritized using Endeavour, leading to the identification of GCKR as the best candidate gene among the 106 genes from the region. They present sequencing evidence that GCKR is a susceptibility gene in Japanese families with clustered diabetes.

4.3.2 Improvement of the text-mining source

The results presented in this section have been obtained through an internal collaboration with Shi Yu, a former PhD student in our group [271, 270].

One important data source is 'Text' and is built by text mining of the scientific literature from the Medline repository. The original 'Text' model was built using one vocabulary derived from Gene Ontology using the Term Frequency times Inverse Document Frequency (TFIDF) representation scheme [4]. It was however still unknown what was the effect of using various vocabularies, representations and ranking algorithms for gene prioritization by text mining. We have systematically investigated this issue by benchmarking five domain vocabularies, two text representation schemes and four classes of ranking algorithms (for a total of 27 algorithms) [271] representing 270 distinct configurations in total. The five domain vocabularies are based on reputed biomedical ontologies: eVOC, MeSH, Gene Ontology (GO), Online Mendelian Inheritance in Man (OMIM) and London Dysmorphology Database (LDDB). For comparison purpose, text mining with no predefined vocabulary is also performed. The two text representations are the

classical Inverse Document Frequency (IDF) and Term Frequency times Inverse Document Frequency (TFIDF) and the four ranking algorithm classes include one-class Support Vector Machine (1-SVM) [210], k-nearest neighbors (KNN), and two clustering based ranking techniques: k-means clustering and hierarchical clustering. First conclusion, the results show that the IDF representation of gene performs better than the TFIDF representation. IDF outperforms TFIDF for all vocabularies and all ranking algorithms. Second conclusion, the eVOC and MeSH domain vocabularies perform better than the other domain vocabularies (GO, OMIM and LDDB). This general conclusion stands for the two vector representations and the best performing algorithms. Last conclusion, the ranking algorithm based on 1-SVM, standard correlation and ward linkage method provides the best performance. A direct consequence of this study was to replace the existing text model of our gene prioritization framework by a model that exhibits a better performance on benchmark data sets.

We have then investigated if the incorporation of more text mining models may be beneficial to obtain more refined and accurate knowledge. To this end, we have developed a multi-view approach to retrieve biomedical knowledge using different controlled vocabularies [270]. These controlled vocabularies were selected on the basis of nine well-known bio-ontologies and were used to index the vast amount of gene-based free-text information available in the MEDLINE repository. Beside the five vocabularies mentioned above, the Kegg Orthology (KO), the Mammalian Phenotype Ontology (MPO), the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT), and the Universal Protein Knowledgebase (UniprotKB) were also considered. The text mining result specified by a vocabulary was considered as a single view and the obtained multiple views were integrated by multi-source learning algorithms (order statistics and one class SVM). We investigated the effect of integration for gene prioritization and systematically evaluated and compared the performance on benchmark data sets. This revealed that the multi-view approach could demonstrate significantly better performance than the other methods. Another finding is that applying Latent Semantic Indexing (LSI) on the gene profiles reduces the dimension of these profiles but also leads to an increase in performance for gene prioritization.

4.3.3 Optimization of the training

One of the key step for an effective gene prioritization is the selection of the training set. Due to the nature of the algorithm, it is important that the training set is homogeneous and represents a single biological process. The presence of genes that do not belong to the pathway / disease described by the other genes represents noise and is likely to have a negative effect on the prioritization results. To estimate this effect, we have performed an additional cross-validation using gene sets that contain a mixture of both disease genes and random genes. The results are displayed in

Percentage of randomly selected genes added	AUC
0%	94,35%
25%	84,83%
50%	78,69%
75%	74,12%
100%	71,43%

Table 4.2: Effect of adding randomly selected genes to the disease training sets. An OMIM benchmark that consists of 29 diseases is used. The amount of randomly selected genes added is proportional to the size of the original disease set, and the AUC is computed accordingly.

Table 4.2. The performance with the original training sets is 94,35%, when half of the sets consist of randomly selected genes, this decreases to 71,43%. Altogether, these results show that adding randomly selected genes to the disease specific gene set potentially reduces the quality of the model and therefore the performance of the associated predictions.

Building such a training set represents a subsequent amount of work, and therefore users would greatly benefit from methods that can automatically or semiautomatically build a training set. The retrieval of known disease and pathway genes from reference databases represents the first step towards this goal, and has been implemented within our framework. It is however often the case that a training set built this way is not homogeneous (making its profiling more difficult) and incomplete. In addition, diseases are often resulting from the perturbation of complex cascades of pathways making their profiling even more challenging. An example is leukaemia, a term used to describe a group of phenotypically similar diseases that, however, differ at the molecular level. To this end, we have developed two complementary strategies: the automatic clustering of the training set and the use are several training set in conjunction to prioritize candidate genes.

Clustering is the repartition of elements from one group into several subgroups so that elements in one subgroup are similar to each other and dissimilar to the elements of the other subgroups. In our case, the clustering of the training set allows the definition of several more homogeneous subgroups. Consequently, it allows the removal of the outliers, i.e., genes that do not belong to any cluster and that therefore represent noise in the data. In a study performed in collaboration with Francisco Bonachela-Capdevila, we were able to show that performing clustering before the prioritization can lead to better cross-validation performance. Clustering was used on binary annotations data sources using the CLOPE algorithm, these data sources were then excluded from the prioritization analysis.

The repartition, manually or through clustering, of the training genes in several

training sets represents the first step. Having multiple training sets means performing multiple prioritizations and thus multiple rankings. To reconcile these results, several methods have been investigated. The first method is the use of an extra layer of order statistics to combine the prioritization results similarly to what is done within Endeavour to fuse results from different data sources. This method assesses that the guilty gene is active in all the biological processes described by the training sets. An example is the prioritization of a region deleted in patients with several phenotypes (associated to the region), and for which the disease causing gene is expected to contribute to all phenotypes. The prioritization of the genes from that region with several training set (one per phenotype) is likely to give better results than if all phenotypes are grouped together in a single training set. In collaboration with Bernard Thienpont, we have implemented and applied this strategy for congenital heart defect (CHD) [232]. A candidate region is defined by genotype-phenotype correlation on chromosome 6q24-q25. This region that contains 105 candidate genes is then prioritized using seven training sets that are all related to CHD (first heart field, second heart field, neural crest, vascularization, left-right asymmetry establishment, valve formation, known dosage-sensitive CHD genes). The seven prioritizations are then combined with the order statistics to create a global ranking. The experimental validation that follows these prioritization has proved that the gene TAB2, ranked first over the locus is associated to CHD, this part is discussed in further details in chapter 8.

An alternative consists in selecting the best results from all training sets for every gene instead of using the order statistics. The underlying assumption is then that the disease gene is involved in one of the biological processes described by all the training sets but not all of them. However, we have not tested this alternative on real data.

Chapter 5

Kernel-based data fusion for gene prioritization

5.1 Summary

The original prioritization strategy presented in chapter 3 is based on basic statistics and every data source has its own modeling and scoring method, which makes it difficult to extend with novel data sources and organisms. It also makes difficult the extension of the approach with new algorithmic developments. The present chapter investigates the development of an alternative strategy that uses more advanced machine learning methods, kernel based methods, to solve the same probelm: candidate gene prioritization through data fusion. In this setup, all the data sources are first transformed into kernels using a linear function or a Radial Basis Function (RBF). Then, a one-class SVM algorithm is applied to perform novelty detection using multiple kernels at the time. A classical one-class SVM algorithm is using a single data source and finds the hyperplane that best separates the positive genes from the origin. In our case, multiple data sources, *i.e.*, multiple kernels, are used in conjunction. We investigate whether the optimal convex optimization of the kernels performs better than the simple average kernel. We also compared the developed method with the approach used in the Endeavour software on the same disease benchmark. The main finding of this study is that kernel methods outperforms the regular Endeavour approach, this is further discussed in section 5.3.

BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–9

Kernel-based data fusion for gene prioritization

Tijl De Bie [a,b], Léon-Charles Tranchevent [c], Liesbeth Van Oeffelen,[c], Yves Moreau,[c]

[a]Dept. of Eng. Math., University of Bristol, [b] OKP Research Group, Katholieke Universiteit Leuven, [c] ESAT-SCD, KULeuven

ABSTRACT

Motivation: Hunting disease genes is a problem of primary importance in biomedical research. Biologists usually approach this problem in two steps: first a set of candidate genes is identified using traditional positional cloning or high-throughput genomics techniques; second, these genes are further investigated and validated in the wet lab, one by one. To speed up discovery and limit the number of costly wet lab experiments, biologist must test the candidate genes starting with the most probable candidates. So far, biologists have relied on literature studies, extensive queries to multiple databases, and hunches about expected properties of the disease gene to determine such an ordering. Recently, we have introduced the data mining tool ENDEAVOUR [1], which performs this task automatically by relying on different genomewide data sources, such as Gene Ontology, literature, microarray, sequence, and more.

Results: In this paper, we present a novel kernel method that operates in the same setting: based on a number of different views on a set of training objects, a prioritization of test objects is obtained. We furthermore provide a thorough learning theoretical analysis of the method's guaranteed performance. Finally, we apply the method to the disease data sets on which ENDEAVOUR [1] has been benchmarked, and report a considerable improvement in empirical performance.

Availability: The MATLAB code used in the empirical results will be made publicly available.

Contact: tijl.debie@gmail.com, yves.moreau@esat.kuleuven.be

1 INTRODUCTION

Identifying genes whose disruption causes congenital or acquired disease in humans is a major goal of genetics and molecular biology, both towards diagnosis and understanding the biology of disease processes. These genes are called *disease genes*—an example being the BRCA1 gene whose mutation is responsible for cases of familial breast cancer. Several biological strategies are available to identify disease genes. Positional cloning strategies aim at identifying the position of the gene on its chromosome (linkage analysis, linkage disequilibrium, association studies, study of chromosomal aberrations). Most of the time these studies can only restrict the location of the disease gene to a region containing tens to hundreds of *candidate genes*. High-throughput genomic studies (microarray analysis, proteomics, and so on) often consider biological samples from patients or animal models and try to identify which key genes or proteins are disrupted in the disease process. Again, these strategies often deliver long laundry lists of hundreds of candidate genes.

In both cases, the candidate genes need to be further investigated to identify the disease causing genes. Because this work is time consuming and expensive, biologists must prioritize the genes from most to least promising when carrying out the validation process—this is called *gene prioritization*.

A main strategy to prioritize candidate genes is to compare the candidate genes (called here the *test genes*) to genes already known to cause the same disease or closely related disease processes (called here the *training genes*). Hence, the problem faced by the biologist to determine the implicated gene among the test genes can potentially be simplified, by concentrating on those test genes that are in some sense similar to the training genes.

With the advent of high-throughput technologies, many sources of information, or *views* on genes may be useful and relevant in defining what is 'similar'. Therefore, this task has become extremely challenging for biologists. For this reason, we have recently developed the tool ENDEAVOUR [1]. It makes use of statistics to compute a ranking of test genes according to their similarity to the training genes, and this once on each of a number of data sources. In a subsequent step, these rankings are integrated into a single ranking by making use of order statistics.

1.1 Formal problem setting

In the current paper, we formulate the problem in machine learning terms, and we develop a kernel-based method to solve it (see [11] for an introduction to kernel methods). As for the formalization, several avenues that can be followed. First, one may cast it into the *classification* framework, regarding the training genes as belonging to the positive class, and the rest of the genome to the negative. However, the assumption that the rest of the genome contains only negatives is false, even though in gene hunting the proportion of positives is usually small. This means that label noise is unavoidable, which is detrimental from a robustness point of view. Moreover, the set of positives is usually extremely small (a few to a couple tens) and is drawn with major biases from the underlying positive class, which compromises uniform generalization performance over the whole space. On the other hand, the large size of the negative training set would pose computational challenges to data fusion approaches.

The second possible approach formalizes the problem as *novelty detection*, where one tries to model (the support of the distribution of) the training genes only. Several approaches to novelty detection have been described in literature [13, 10], and relations between them have been established. One approach tightly fits a hypersphere around a vector representation of the data, and considers the inner volume of the hypersphere as the support of the distribution. Another approach finds a hyperplane separating the positive data from the origin.

The approach proposed in this paper is reminiscent mostly of the latter: find a hyperplane that separates the vector representations of

the disease genes from the origin with the largest possible margin, and consider a gene more likely to be a disease gene if it lies farther in the direction of this hyperplane. However, we have an additional problem to be dealt with: while all methods described so far make use of just one view on the data, our method should be capable of taking into account several different views on the genes.

1.2 Data fusion by learning the kernel

The main source of inspiration for our algorithmic and theoretical contributions is in our previous work [5], and in [7, 6, 3]. They describe a methodology for learning the kernel matrix relying on a quadratically constrained linear program (QCLP) for classification in a transduction setting. Both approaches rely on strong statistical foundations and performance guarantees are provided. A number of recent publications have carried this work further, generalizing this approach towards other problems besides classification [9], working on algorithmic improvements to reduce time and memory requirements [2], or contributing to both these aspects [12]. Still, thus far data fusion approaches to novelty detection have remained understudied. To our knowledge, a statistical study of the problem is still lacking. Furthermore, empirical studies have remained limited to the method in [9], which is based on the elegant framework of kernel-learning with hyperkernels, but which is computationally extremely challenging as it relies on a semi-definite program with a number of variables that is quadratic in the training set size.

1.3 Results

We present an approach for gene prioritization based on a novel kernel-based algorithm capable of integrating various sources of information in a natural way. Our approach leads to fast algorithms relying on a QCLP, and we show how it is explicitly guided by a rigorous statistical study. We demonstrate it on a large number of disease gene hunting problems, outperforming ENDEAVOUR [1], which is the first tool to combine many data sources for gene prioritization and to have provided new candidate genes that have been successfully biologically validated.

2 DATA FUSION, KERNEL COMBINATIONS, AND NOVELTY DETECTION

We first discuss a variant of a well-known kernel method for novelty detection, which makes use of a single view on the data only [10]. Let us assume a gene x has an associated vector representation $\mathbf{x}$. Then, this method finds a hyperplane parameterized by a unit norm weight vector $\mathbf{w}$, defined by the equality $f(x) \triangleq \mathbf{x}'\mathbf{w} = M$, which separates all training genes from the origin.

Then, for a *test* gene x, the function f measures its distance from the origin in the direction of the hyperplane, and can be used to prioritize the genes: the larger $f(x)$, the higher gene x in the prioritization. See the right part of Figure 1 for a schematic clarification.

Subsequently, we discuss how different views on the genes can be integrated naturally and efficiently by convexly combining the kernels of each of these data sources. The statistical study which theoretically supports the use of the function f as a way to prioritize is left to Section 3.

2.1 Novelty detection

Let us represent the vector representation of the set of training genes by a matrix $\mathbf{X}$, with the ith row of $\mathbf{X}$ containing the feature vector $\mathbf{x}_i$ of the ith training gene. As above, we define a function f of gene x as $f(x) \triangleq \mathbf{x}'\mathbf{w}$. Then, we search for the weight vector $\mathbf{w}$ with $\|\mathbf{w}\|^2 \leq 1$ such that for all genes x_i in the training set the function $f(x_i)$ is larger than a margin M, with M as large as possible (i.e., it searches for a hyperplane parameterized by $\mathbf{w}$, such that all training data lie on one side of the hyperplane and the perpendicular distance M between the origin and the hyperplane is maximized). Formally, this leads to the optimization problem,

$$\max_{M,\mathbf{w}} p(M) = M \quad \text{s.t.} \quad \mathbf{w}'\mathbf{w} \leq 1, \;\; f(x_i) \triangleq \mathbf{x}_i'\mathbf{w} \geq M \; (\forall i).$$

The dual of this (convex) optimization problem can be given by

$$\min_{\boldsymbol{\alpha}} d(\boldsymbol{\alpha}) = \sqrt{\boldsymbol{\alpha}'\mathbf{X}\mathbf{X}'\boldsymbol{\alpha}} \quad \text{s.t.} \quad \alpha_i \geq 0 \; (\forall i), \;\; \mathbf{1}'\boldsymbol{\alpha} = 1. \; (1)$$

Thanks to strong duality, the primal and dual optima (achieved for $M^\star$ and $\boldsymbol{\alpha}^\star$) are equal to each other: $p(M^\star) = d(\boldsymbol{\alpha}^\star)$. Furthermore, duality relations show that the optimal value of the weight vector can be expressed in terms of the dual variables as $\mathbf{w}^\star = \mathbf{X}'\boldsymbol{\alpha}^\star / \sqrt{\boldsymbol{\alpha}^{\star\prime}\mathbf{X}\mathbf{X}'\boldsymbol{\alpha}^\star}$. Note that the square root is a monotonic function and can hence be ignored in the objective of the dual optimization problem (1).

It is a crucial recurring fact in kernel methods that the dual formulation can be written solely in terms of inner products between feature vectors $\mathbf{x}_i$. Indeed, the matrix $\mathbf{X}\mathbf{X}'$ contains the inner product $\mathbf{x}_i'\mathbf{x}_j$ on its ith row and jth column, and we denote $\mathbf{X}\mathbf{X}' = \mathbf{K}$, the so-called kernel matrix. As a consequence, the actual representation $\mathbf{x}$ of data object x does not need to be known, as long as the inner product between any pair of objects in this representation is specified by a kernel function $k(x_i, x_j)$.

Equally importantly, instead of the representation $\mathbf{x}'\mathbf{w}$ of the prioritization function $f(x)$, we can use the following equivalent dual formulation, relying on kernel evaluations only:

$$f(x) = \frac{1}{\sqrt{\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}}} \sum_{i=1}^{n} \alpha_i k(x, x_i) \quad (2)$$

2.2 Data fusion

In the method discussed above only a single kernel function k and corresponding kernel matrix $\mathbf{K}$ on the training set are given. However, in our application m different kernel functions k_j and kernel matrices $\mathbf{K}_j$ are available, each of which is based on a certain representation or view on the genes. The availability of these different views leaves us the freedom to design the kernel matrix such that the penalized margin is maximized. The challenge is to exploit this in a statistically and algorithmically sound way.

2.2.1 Averaging the kernels A first and trivial approach is to combine the kernel matrices $\mathbf{K}_j$ by simply computing a weighted average,

$$\mathbf{K} = \sum_j \frac{\mathbf{K}_j}{\beta_j},$$

with $\beta_j > 0$ positive constants. The choice of the constants β_j may be arbitrary and at the user's discretion. However, here we suggest a simple agnostic choice. The concern to address using these weights is that different kernels may have different scales, such that their importance in the linear combination may be overly small or

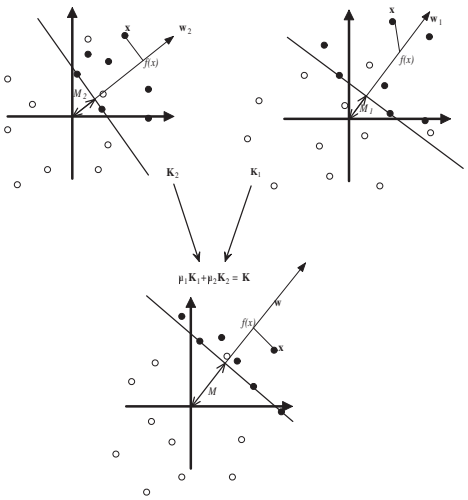

Fig. 1. Schematic representation of the hyperplane separating the (positive) training genes (filled circles) from the origin, along with the negative genes (open circles). Combining two kernels in an optimal way leads to a new space (right figure) where the distance of the positive genes to the origin is larger.

large. In order to correct for this, we choose β_j to be proportional to the trace of $\mathbf{K}_j$. Essentially, this value for β_j is chosen in order to make the kernels comparable to each other. Statistical arguments for this choice will be given in Section 3. This is the first approach we propose and compare in the experimental Section. There we will see that it outperforms ENDEAVOUR significantly and by a large margin.

Note that, if appropriate, one could hand-tune the weights of the kernels in the kernel combination based on expert knowledge. I.e., instead of using $\mathbf{K}$ as defined above, one can additionally weigh the kernels with a hand-tuned weight $\mu_j \geq 0$:

$$\mathbf{K} \quad = \quad \sum_j \mu_j \frac{\mathbf{K}_j}{\beta_j}$$

A kernel for a data source deemed relevant for a certain disease could then be given a larger μ_j, and hence a larger vote in the linear combination. Since this requires expert knowledge on individual diseases, we choose not to follow this avenue in the present paper.

Nevertheless, as we will see below, there are other ways to tune the weights μ_j automatically in a data-dependent but agnostic way (i.e., without taking disease characteristics into account). The goal of these approaches is to reduce the influence of noisy (or irrelevant) information sources, and of double counting of information that is present in more than one of the information sources. Hence, as we will see in the experimental Section, in cases where noise influences are large or where redundant information is provided, such methods may perform better. Let us discuss these methods now in greater detail.

2.2.2 Optimal convex kernel combination As a first method to achieve automatic tuning of the kernel weights, and in the same spirit as [7, 5], we propose to convexly combine the kernel matrices $\mathbf{K}_j$ so as to maximize the margin M between the data points and the origin. More specifically, with β_j positive constants, we choose the 'summarizing kernel' $\mathbf{K}$ as the one from the set $\mathcal{K} = \left\{\sum_j \mu_j(k_j/\beta_j) : \boldsymbol{\mu}'\mathbf{1} = m, \boldsymbol{\mu} \geq \mathbf{0}\right\}$ that maximizes the optimum of optimization problem (1):

$$\max_{\mathbf{K}} \min_{\boldsymbol{\alpha}} \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} \quad \text{s.t.} \quad \alpha_i \geq 0\ (\forall i),\ \mathbf{1}'\boldsymbol{\alpha} = 1,$$
$$\mathbf{K} \in \left\{\sum_j \mu_j \frac{\mathbf{K}_j}{\beta_j} : \boldsymbol{\mu}'\mathbf{1} = m, \boldsymbol{\mu} \geq \mathbf{0}\right\}.$$

which can be shown to be equivalent to

$$\min_{t,\boldsymbol{\alpha}} t \quad \text{s.t.} \quad \alpha_i \geq 0\ (\forall i),\ \mathbf{1}'\boldsymbol{\alpha} = 1, \qquad (3)$$
$$t \geq \boldsymbol{\alpha}'\frac{\mathbf{K}_j}{\beta_j}\boldsymbol{\alpha}\ (\forall j).$$

This is a QCLP problem that is efficiently solvable using general purpose software.

2.2.3 A regularized intermediate solution In some cases, the freedom allowed to the optimization problem in this way may be so large that overfitting occurs, resulting in a bad generalization performance. Therefore, we propose an approach intermediate to the simple averaging of the kernels and their optimal combination using convex optimization as explained above. This can be achieved by specifying a lower bound $0 < \mu_{\min} \leq 1$ on μ_j in the specification of $\mathcal{K}$, i.e. $\mathcal{K} = \left\{\sum_j \mu_j(k_j/\beta_j) : \boldsymbol{\mu}'\mathbf{1} = m, \boldsymbol{\mu} \geq \mathbf{1}\mu_{\min}\right\}$. Increasing $\mu_{\min}$ reduces the size of $\mathcal{K}$, which amounts to regularizing the problem, and hence reduces the risk of overfitting. For $\mu_{\min} = 1$, the simple method that computes a weighted average of the kernels is obtained.

2.2.4 A unifying method Interestingly, the last method contains the first and the second as a special case. Indeed, by taking $\mu_{\min} = 1$, the first method is obtained. By taking $\mu_{\min} = 0$, the second method is obtained. Therefore, in the remainder of the paper, we can refer to the different methods by choosing the value for $\mu_{\min}$.

2.2.5 The function f For all these data fusion approaches, the evaluation of the function $f(x) = \mathbf{x}'\mathbf{w}$ can now be expressed in terms of the β_j and α_i, as

$$f(x) = \frac{1}{\sqrt{\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}}} \sum_{i=1}^{n} \alpha_i \quad \sum_{j=1}^{k} \mu_j \frac{k_j(x, x_i)}{\beta_j}\Bigg). \qquad (4)$$

In the next Section we will provide a rigorous theoretical evidence motivating these approaches, which will furthermore point us to good possible choices for values of β_j.

3 STATISTICAL GUARANTEES AND MOTIVATIONS

While earlier approaches to novelty detection based on different kernel-views exist (e.g. [9, 12]), very few experimental results have been reported (only in [9], and only in an informal qualitative way). Furthermore, to our knowledge, none of these were based on statistical foundations. Still, it is clear that in our above formulation, certain parameters need to be chosen, in particular the values of β_j. Additionally, we will see that other choices need to be made, regarding normalizations and centering of the data. In order to make these choices in a principled way, a statistical study is indispensable. Here we present such a study, which, as in [7], relies on the use of Rademacher complexities.

In our discussion, we focus on the unregularized version, where $\mu_{\min} = 0$. All results can quite easily be adapted to the regularized case, and we will point out consequences of this regularization where relevant.

3.1 Controlling the number of false negatives

We will assume that the training genes are sampled *iid* from the distribution of the positive class (the class of disease genes). Admittedly, this is not exactly true, but it is probably a good approximation given the large number of genes in the genome. We derive a bound for the probability that $f(x) \leq M - \gamma$ for an *iid* test gene x from the positive class. This is the probability to make an error of a certain magnitude in evaluating whether the test point x is a novelty or not, i.e., the probability that a test point from the positive distribution lies a distance γ at the negative side of the hyperplane. We will see that this probability quickly decreases for increasing γ, which means that the probability that for a true disease gene x the function $f(x)$ can be expected to be large. Let us now state the Theorem; for brevity, we provide its proof in Appendix.

THEOREM 1. *Given a set X of n objects (genes) x_i sampled* iid *from an unknown distribution $\mathcal{D}$. Let $\lambda(\mathbf{K}_j)$ denote the largest eigenvalue of $\mathbf{K}_j$. Then, for any $M, \gamma \in \Re_+$ and for any $\delta \in (0,1)$, with probability of at least $1-\delta$ the following holds for the function f in Equation (4) as found by optimization problem (3):*

$$P_{\mathcal{D}}\left(f(x) \leq M - \gamma\right) \leq \frac{1}{n\gamma} C(\mathbf{K}_j, \beta_j) + \frac{1}{\sqrt{n}} \sqrt{2 \ln \frac{2}{\delta}}.$$

where $C(\mathbf{K}_j, \beta_j)$ is a complexity term equal to

$$C(\mathbf{K}_j, \beta_j) = 4 \sqrt{\min \quad n \max_j \frac{\lambda(\mathbf{K}_j)}{\beta_j}, \sum_{j=1}^{m} \frac{\operatorname{trace}(\mathbf{K}_j)}{\beta_j}}\Bigg).$$

While the Theorem holds for any specific value of M, it does not hold uniformly over M. However, as outlined in [7] and references therein, the theorem can easily be adapted to yield uniform bounds over M. Note that the complexity term involves the square root of the minimum of two quantities, the first of which grows quadratically with n, while the second grows only proportionally with n as the training set grows. Hence, asymptotically, the bound decreases as $1/\sqrt{n}$.

A similar Theorem holds for the regularized version, where a lower bound is imposed on the values of μ_j. Interestingly, in that case the bound is generally tighter, as the Rademacher complexity of the smaller function class is smaller. We omit the Theorem here for readability.

In some sense, Theorem 1 bounds the false negative probability by a value that depends on the value of β_j relative to $\lambda(\mathbf{K}_j)$ and to $\operatorname{trace}(\mathbf{K}_j)$, which should therefore be kept under control. We will now show how an approach to controlling the false positives motivates a choice for the β_j that ensures this requirement is fulfilled in practical situations.[1]

3.2 Controlling the number of false positives

Given the full genome, and the probability of false negatives being bounded, we could control the number of false positives by bounding the total number of positives. Recall that the algorithm tries to separate the data as far from the origin as possible. With this in mind, we suggest the following strategy. First, use centered kernel matrices (or kernel functions), i.e. in gene hunting the kernels are defined by centering the kernels on the full genome. And second, equate the value of β_j to the trace of the jth centered *genome-wide* kernel matrix divided by the total number of genes in the genome. Such a choice for β_j can be expected to yield tight capacity terms in Theorem 1 in practice, assuming that positive and negative genes are not *too* different in norm and in distribution (a reasonable assumption, as it appears so in practice and it is the facts motivating this work).

This strategy ensures that the trace of the kernel matrix obtained by linearly combining all genome-wide kernel matrices weighted by β_j has a trace equal to the number of genes in the genome, such that the norm $k(x,x)$ of a gene is equal to 1 on average. Hence, for a function $f(x)$ as found by any of our 3 data fusion methods, the centering and choice of β_j imply that $E_x(f(x)) = 0$ and $E_x(f(x)^2) = (\mathbf{w}'\mathbf{x})^2 \leq \|\mathbf{w}\|^2 \|x\|^2 = \|\mathbf{w}\|^2 k(x,x) \leq 1$. In this way, for a large margin between the training points and the origin, one can expect that relatively few data points lie at the positive side of the hyperplane, as quantified by e.g. the one-tailed Chebyshev's inequality:

$$P \quad f(x) - E_x f(x) \geq k\sqrt{E_x(f(x)^2)}) \quad \leq \frac{1}{1+k^2},$$

where the expectations are over the total gene distribution. Applied to our problem:

$$P(f(x) \geq k) \leq \frac{1}{1+k^2}.$$

In practice, the number of genes in the genome is so large that *iid* assumptions concerning test and training genes become realistic. In such cases it is possible (and more convenient) to carry out the kernel-centering and determination of β_j on a smaller number of genes, such as on the combination of test and training-genes. This is the approach we take in the experiments below.

[1] In this context we would like to note that, while the theory and the algorithm for transduction in [7, 5] is never explicitly expressed in terms of such β_j that weigh the kernel matrices, also there a similar choice has been made. In that paper, what would be the equivalent of our β_j is chosen to be equal (or proportional) to the trace of the kernel matrix on the training and test points together. This choice is also implicitly motivated by the statistical study.

4 GENE PRIORITIZATION RESULTS

We will now apply our algorithms on a series of actual biological data on disease gene hunting, taken from a large-scale cross-validation study from [1]. In this paper, 29 diseases are investigated, and to each disease a number of genes between 4 and 113 are known to be associated, with 624 as the total number of disease genes in the study. To assess the performance of a gene-hunting method the following strategy is used. Note that for comparability, we choose to use an essentially identical assessment strategy as the one used in [1]. *For each disease*, do the following:

1. Choose a set of 99 genes, randomly selected from the genome.
2. Perform leave-one-out cross-validation: Regard one of the training genes as a test gene. This test gene is further referred to as the hold-out gene. Then apply the disease gene hunting method to the reduced training set which is obtained by omitting the hold-out gene. Ideally, the hold-out gene will be unveiled as a disease gene, which means that it will be on top of the list. To verify this, record the rank of the hold out gene in the test set. Since there are 99 test genes and 1 hold-out gene hidden among them, this rank will be in between 1 and 100.

Now, based on all these ranks in the cross-validation (over all diseases and all disease genes for these diseases), construct a ROC-like curve in the following way. Plot the fraction of genes that, when held out, rank among the top $x\%$ of test genes, and this as a function of x. If in each hold out experiment the hold out gene ranks first, the ROC-like curve will be 1 for all x, and the area under the ROC curve, further called AUC (Area Under Curve) is equal to 1. For a random training and test set combination, the AUC is 0.5 in expectation.

4.1 Data sources and kernels used

4.1.1 The data sources We are using the following data sources, based on Ensembl v39[4]: microarray data (MA), DNA sequence (Seq), EST data (EST), Gene Ontology annotations (GO), InterPro domains (IP), KEGG pathways (KEGG), motifs (Motif), binding data (BIND), and literature (Text), just as in [1]. Not to obfuscate the comparison, in our method we deal with missing values in the most naive way, e.g. by equating them to genome-wide averages. We should note that in the ENDEAVOUR paper [1] data from a previous version of Ensembl was used. Therefore, to ensure a fair comparison, we reran their experiments with ENDEAVOUR based on the most recent Ensembl version v39 as well. Hence, we will compare our proposed methods with the performance of ENDEAVOUR on exactly the same data.

4.1.2 The kernel matrices For each data source except for Seq, we use three different kernels:

1. the linear kernel followed by normalization,
2. a Radial Basis Function (RBF) kernels with kernel width equal to twice the average distance of a data point to its nearest neighbor in the union of the training and the test set,
3. and an RBF kernel with kernel width equal to four times the average distance of a data point to its nearest neighbor in the complete data.

The kernel widths are chosen heuristically according to rules of thumb that often yields good results in practice. For the sequence data, we also used three different kernels: the 2-mer, 3-mer, and 4-mer kernels as defined in [8]. Hence, in total 27 kernels are used, 3 for each of the 9 data sources.

4.1.3 Noise data sources As explained below, we have also carried out a robustness analysis by constructing random noise data sources to be included as additional data. These noise data sources consist of 10-dimensional normally distributed random vectors (variance equal to 1). For each noise data source, we constructed 3 kernels to use in the algorithm, a linear one and two RBF kernels, exactly as for the other vectorial data sources. Constructed in this way, a noise data sources should quite accurately mimic real-life data with no relevance to the problem. Note that a comparison with ENDEAVOUR in terms of noise robustness is hard to design, since true noise models cannot as easily be generated as we can generate noise kernels. Therefore, we will exclude ENDEAVOUR from the noise robustness analyses below.

4.2 Disease genes hunting: results

We have carried out a number of experiments to assess the following:

1. the performance gain when compared to ENDEAVOUR of the simple method with uniformly weighted kernels ($\mu_{\min} = 1$),
2. the use of automatically tuned weights when noisy data sources are taken into account, or when a small number of data sources is much more informative than the others ($\mu_{\min} > 0$),
3. the use of automatically tuned weights with a lower bound in the same scenario ($\mu_{\min} \geq \mu_{\min} = 0.5$).

In order to assess noise resilience, we examined the performance as a function of the number of noisy data sources, ranging from 4, over 8, to 16 noise sources, yielding 12, 24, and 48 kernels respectively.

Lastly, we performed each of these same experiments in three scenarios: (i) based on all data sources listed above, (ii) based on all but Text, (iii) based on all but Text and GO, and (iv) based on all but Text, GO and KEGG. We have performed these exclusions in order to investigate to what extent the methods are capable of extracting information from data that may lead to novel discoveries, as opposed to for example Text data that may contain known clues of disease implications. We will now discuss the results in detail.

In order to obtain stable results, we performed 10 randomization for each experiment reported below. In each of these randomizations, a different set of test genes has been chosen, randomly selected from the genome. The same random test set was used in the different methods being compared.

4.2.1 Comparison of the uniformly weigthed method with ENDEAVOUR We compared the performance of ENDEAVOUR with our method with $\mu_{\min} = 1$. The results are summarized in Table 1, and clearly show that the proposed method outperforms ENDEAVOUR significantly, and by a large margin. This is the case for all (sub)sets of data sources investigated.

Furthermore, we should note that the proposed method is computationally extremely fast: finding the optimal $\boldsymbol{\alpha}$ takes a negligible time for up to 100s of training genes, and computing the ranking function f on a test gene (the testing phase) is extremely fast as well.

Table 1. Comparison of the simple, uniformly weighted, kernel method with ENDEAVOUR. Different scenarios are considered, taking into account all 9 data sources, all but Text, all but Text and GO, and all but TEXT, GO, and KEGG. The first two lines show 1-AUC (lower is better) for both methods, averaged over 10 random selections of the test genes. The last line shows a p-value computed by means of a paired t-test, testing the null hypothesis that the expected 1-AUC is not smaller for the kernel method than for ENDEAVOUR. Clearly, the difference is highly significant for all but the last set of data sources used. Furthermore, the AUC value differs considerably for the first 3 scenarios considered.

1-AUC	All	No Text	No Text, GO	No Text, GO, KEGG
ENDEAVOUR	0.0833	0.1290	0.1698	0.1698
Kernel method	**0.0686**	**0.1043**	**0.1491**	**0.1675**
p-value	7.4e-10	7.5e-11	3.3e-7	2.4e-1

4.2.2 Performance in the presence of one or few dominantly informative information source It can be assumed (and it is observed) that in general, Text and GO contain the most accessible information relevant to disease gene hunting. While this may not always be the case, it is possible that in other cases another data source contains is much more relevant than any of the others. Therefore, it is of interest to assess to what extent the different methods are able to disregard the less informative data sources.

To assess this, in Table 2 we summarized the 1-AUC scores for different subsets of data sources, and this for our proposed methods and for ENDEAVOUR. We can conclude that if there is a clear best information source (e.g. Text or GO), it pays off to tune the weights automatically (either with $\mu_{\min} = 0$ or, better, with $\mu_{\min} = 0.5$). If all or most information sources are roughly equally good, the uniform weighting performs better than with the automatically tuned weights. In all cases except when all data sources including Text are being used, $\mu_{\min} = 1$ seems to perform comparably well to the other kernel methods, and in all cases it performs better than ENDEAVOUR (see higher).

4.2.3 Investigation of the noise sensitivity Table 2 reveals that for increasing amounts of noise, the uniformly weighted method degrades much more rapidly than the methods with tuned weights. This can be explained by the fact that the tuned weights are usually lower for noise kernels than for the informative kernels. Similarly, a larger number of (approximate) copies of the same bad kernel would degrade the performance of the naive method with equal weights, while the methods with tuned weights are insensitive to this. For a discussion of similar observations in a related context, see [7, 5].

Overall, when large amounts of noise are to be expected or when one or few of the data sources is much more discriminative than the others, the regularized method ($\mu_{\min} = 0.5$) seems the most robust and performant method. If less than half of the information sources are suspected to be irrelevant, it is better to use the uniformly weighted kernel method ($\mu_{\min} = 1$).

4.2.4 Performance of individual kernels versus the overall performance Besides comparing the different proposed data fusion methods, we should assess whether it makes sense at all to perform data fusion. To this end, consider Figure 2. We now only consider the regularized method with $\mu_{\min} = 0.5$, which seems to be a safe approach, though the method with $\mu_{\min} = 1$ performs even slightly better in most noiseless cases we investigated. Nevertheless, it is interesting to investigate the weights attributed to the individual kernels by the method with $\mu_{\min} = 0.5$. Our findings are:

1. When using all data sources, the performance is not significantly different from the performance based on single-kernel novelty detection on Text, in particular for the linear kernel. The most likely reason is that the genes in this study have been well-described, such that Text is likely to be most informative by far. Then, taking other less informative (or more 'noisy') data into account may be expected to worsen the result. However, it is encouraging that this is not the case. (See Figure 2 above.) One may argue that it is as good to simply pick the Text data source (if it is available), and discard the others. It should be noted however that in general it is not known a priori which data source is clearly the better. Hence, also the task to pick the best kernel has to be made in a data drive way, and imperfections in this selection process would degrade the result. Hence, comparing the data fusion methods with any of the single kernels is unfavorable for the data fusion methods.
2. When applied to all data sources except the suspectedly richer ones such as Text, GO and KEGG, our method based on kernel combinations is clearly better than any of the separate kernels. This effect is stronger if more informative data sources are excluded.
3. The weights μ_i attributed to each of the kernels are summarized in the bottom four bar plots of Figure 2. Clearly, when taken into account, Text, and to a lesser extent GO data, get the largest part of the weight. When Text data, or Text and GO, are absent the weights are more evenly distributed. Overall, linear kernels yield better results and get higher weights on this data set. The Seq data gets small weights all over, despite its good individual performance. A potential explanation is that interesting aspects of the sequence information are contained in other information sources, such as InterPro.

In summary, our method effectively integrates complementary information from different sources.

Table 2. The mean AUC performances, for all 3 methods: automatic tuning with $\mu_{\text{min}} = 0$, regularized automatic tuning with $\mu_{\text{min}} = 0.5$, and uniformly weighted (equivalent with using $\mu_{\text{min}} = 1$). ENDEAVOUR's performance is shown for comparison. Clearly, overall the simple kernel-based method with $\mu_{\text{min}} = 1$ comes out best, despite a slightly lower performance than the methods with $\mu_{\text{min}} = 0$ and $\mu_{\text{min}} = 0.5$ for when the Text data is included.

	$\mu_{\text{min}} = 0$	$\mu_{\text{min}} = 0.5$	$\mu_{\text{min}} = 1$	ENDEAVOUR
All data	0.0505	**0.0477**	0.0686	0.0833
No Text	0.1241	0.1121	**0.1043**	0.1290
No Text/GO	0.1902	0.1644	**0.1491**	0.1698
No Text/GO/KEGG	0.2121	0.1828	**0.1675**	0.1698

Table 3. The mean AUC performances, for all 3 methods: automatic tuning with $\mu_{\text{min}} = 0$, regularized automatic tuning with $\mu_{\text{min}} = 0.5$, and uniformly weighted (equivalent with using $\mu_{\text{min}} = 1$), and this for varying number of noise sources. Clearly, the method with $\mu_{\text{min}} = 0.5$ is the most robust against noise. However, also with $\mu_{\text{min}} = 1$ a good robustness is achieved.

		$\mu_{\text{min}} = 0$	$\mu_{\text{min}} = 0.5$	$\mu_{\text{min}} = 1$
All data	No noise	0.0505	**0.0477**	0.0686
sources	$4\times$ noise	0.0596	**0.0579**	0.0950
	$8\times$ noise	0.0656	**0.0644**	0.1144
	$16\times$ noise	0.0702	**0.0694**	0.1420
No Text	No noise	0.1241	0.1121	**0.1043**
	$4\times$ noise	0.1411	**0.1330**	0.1395
	$8\times$ noise	0.1520	**0.1444**	0.1629
	$16\times$ noise	0.1624	**0.1566**	0.1943
No Text,	No noise	0.1902	0.1644	**0.1491**
no GO	$4\times$ noise	0.2186	0.2034	**0.2005**
	$8\times$ noise	0.2375	**0.2257**	0.2275
	$16\times$ noise	0.2554	**0.2496**	0.2599
No Text,	No noise	0.2121	0.1828	**0.1675**
no GO,	$4\times$ noise	0.2410	**0.2245**	0.2296
no KEGG	$8\times$ noise	0.2626	**0.2500**	0.2612
	$16\times$ noise	0.2825	**0.2770**	0.2963

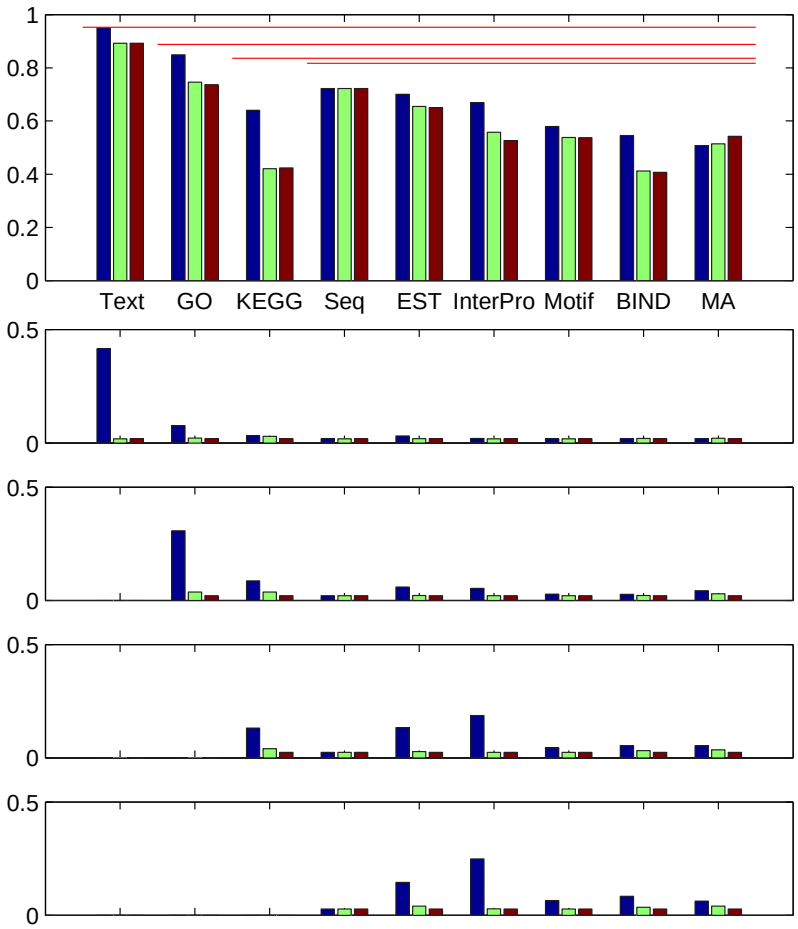

Fig. 2. The top figure shows the AUC performances of each individual kernel, three for each data source (in each triplet the left bar corresponds with the linear kernel, the middle one with the RBF with small kernel width, the right one with large kernel width). For comparison, the 4 full red lines indicate the performance of the kernel combination method with $\mu_{\min} = 0.5$, using all data sources, all but Text, all but Text and GO, and all but Text, GO and KEGG (in order of decreasing AUC). The lower 4 bar plots show the weights μ_j attributed to each of the kernels by the kernel combination method with $\mu_{\min} = 0.5$, using each of these 4 (sub)sets of data sources.

5 CONCLUSIONS AND OUTLOOK

We have presented a new approach and its theoretical analysis, to provide an adequate answer to a recently identified highly relevant problem in bioinformatics. All presented data fusion methods following this approach are shown empirically to outperform the method that has currently been most successful in this setting. Two aspects contribute to its success. First, the kernel method on itself seems to ensure a good performance in this context. Second, a uniform linear combination of all kernel matrices appears to be a robust and highly performant method for data fusion for disease gene hunting. And third, the data-dependent automatic weighting procedures ensure robustness against irrelevant or too noisy data sources.

A particularly appealing aspect of the method is its computational efficiency. The kernels can be computed offline, which makes their computation time less relevant. All other tasks, training is extremely fast (with relatively small training sets of up to a few hundreds), and even more so prioritizing genes can be carried out extremely efficiently, easily scalable to a genome-wide scale.

As further work, we plan to investigate whether a hand-tuning of the weights is a feasible and useable approach in practice. The main question to be answered here is whether the optimal values for the kernel weights represent some intuitive notion of relevance of the kernels.

ACKNOWLEDGMENT

This work is supported by the CoE EF/05/007 SymBioSys, and project GOA/2005/04, both from the Research Council K.U.Leuven.

REFERENCES

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

[2]F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML04)*, 2004.

[3]D. Herrmann and O. Bousquet. On the complexity of learning the kernel matrix. In *Advances in Neural Information Processing Systems 15 (NIPS02)*, 2003.

[4]T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X.M. Fernandez-Suarez, J. Gilbert, M. Hammond, H. Herrero, J. abd Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and Birney E. Ensembl 2005. *Nucleic Acids Res.*, 1(33):D447–D453, Jan. 2005.

[5]G. Lanckriet, T. De Bie, N. Cristianini, M. Jordan, and W. Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[6]G. Lanckriet, M. Deng, N. Cristianini, M. Jordan, and W. Stafford Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputing (PSB04)*, 2004.

[7]G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[8]C. Leslie and R. Kuang. Fast kernels for inexact string matching. In *Conference on Learning Theory and Kernel Workshop (COLT03)*, pages 114–128, 2003.

[9]C. S. Ong, A. Smola, and R. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1043–1071, 2005.

[10]B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[11]J. Shawe-Taylor and N. Cristianini. *Kernel methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K., 2004.

[12]S. Sonnenburg, G. Rätsch, and C. Schäfer. A general and efficient multiple kernel learning algorithm. In *Advances in Neural Information Processing Systems 18 (NIPS05)*, 2006.

[13]D.M.J. Tax and R.P.W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11-13):1191–1199, 1999.

5.2 Contribution of the PhD candidate

The PhD candidate has gathered the genomic data sources used to build the kernels. He has also prepared the benchmark analysis and has performed the Endeavour benchmark to compare with the kernel based method. He has participated in the writing of the gene prioritization section of the paper.

5.3 Discussion

Following the results described in the paper that show an improvement in performance, it was decided to develop a software tool implementing the approach and to make it publicly available. The algorithm described in this paper is implemented in Matlab, which was chosen for its ability to handle fast matrix based calculations and its compatibility with the optimization toolbox SeDuMi. It is however only commercially available, so a different implementation was chosen and led to a tool termed 'MerKator'. This work is described in chapter 6.

A recurring discussion in machine learning is whether the gain in performance obtained by optimizing the weights of the different kernels worths the computing time and power spent for this convex problem. This is still very much an open question, and the results presented in this chapter show that the difference exists but is still small (error of AUC is 0.0686 using uniformly weighted kernel and 0.0505 for automatically optimized fusion). This confirms previously reported results that using the uniformly weighted kernel leads in general to results similar to the ones obtained via the complex optimization [158, 130]. This issue is further discussed in chapter 9.

5.3.1 Improved SVM modeling

We have defined the gene prioritization problem as a novelty detection problem using one-class SVM, thus relying only on the presence of positive genes (the disease causing genes). We motivate our choice by the fact that gene prioritization is not a classification task since the true negative genes are unknown. Several studies propose an approach in which negative sets are approximated by randomly building gene sets from the whole genome. This approach together with repetition, to reduce noise, is used to predict novel transcription factors targets for instance [162]. The authors have observed that using a one class SVM method for the same problem leads to worse performance.

Another recurring problem with the algorithm implemented in the paper is the sparsity of the solution. That is in a majority of cases, the convex optimization

procedure results in one or a few data sources to be favored (high weight), while the other data sources contribute only modestly (*e.g.*, very low weight).

Regarding the disease based benchmark, the most often favored data sources is 'Text'. However, when removing it, the most favored source becomes 'Gene Ontology', without decreasing the performance too much. When both data sources are removed, the weights are distributed to other data sources, again without significant decrease in performance. Furthermore, in some cases, it is possible to obtain better results by removing the data sources that get the higher weight. This shows that the sparseness of the solution does not reflect the relevance of the data sources regarding the benchmark but rather represents an artifact introduced by the algorithm. In the paper, a first solution is introduced by way of adding an additional constraint to the problem that defines a minimal boundary on the weights. This constraint results in a decreased overall performance although it stays higher than the classical approach and also higher than any of the other data sources when used individually. This minimal boundary can be set to 1 so that all data sources contribute equally, which means that the uniformly weighted kernel is used. This approach has however several shortcomings: setting a minimal boundary might not always be a good idea if some of the data sources are indeed noisy or useless regarding the problem under consideration. In this case, a sparse solution is still optimal to avoid predictions based on noisy / useless data. Also, the minimal boundaries are set to be the same for all the data sources. Therefore the majority of the weights that would have been zeros without this additional constraint are still assigned the same value (that corresponds to this minimal boundary). At the end, the solution is not sparse but still no distinction is made between the data sources except for the ones that get the higher weights. A solution obtained with minimal boundary is treating the different data sources equally, which might be suboptimal in some cases.

In a follow-up paper, the effect of optimizing different norms in the dual problem is investigated for gene prioritization, clustering and classification[269]. In optimization, the dual problem is complementary to the primal problem and, in fact, solving one of them solves both. Most of the existing Multiple Kernel Learning (MKL) methods are based on the formulation proposed by Lanckriet *et al.* [130], which can be defined as the optimization of the infinity norm (L_∞) of kernel fusion. Optimizing L_∞ MKL in the dual problem corresponds to posing L_1 regularization on the kernel coefficients in the primal problem, which results in sparseness of the kernel coefficients. Thus, the solution obtained by L_∞ MKL is also sparse, with dominant coefficients to only one or two kernels. At contrary, the L_2 MKL yields a non-sparse solution, which smoothly distributes the coefficients on multiple kernels and, at the same time, leverages the effects of kernels in the objective optimization. Benchmark results show that the L_2-norm kernel fusion leads to a better performance in gene prioritization by data fusion. The figure 5.1 contains the results on a disease benchmark. The L_2 based method shows

the best performance, ahead of the L_∞ based method, showing that it reflects better the relevance of the kernels. The performances of L_2 and the regularized L_∞ are comparable. However, the minimal boundary of the regularized L_∞ is usually predefined empirically while the main advantage of the L_2 approach is that boundaries are kernel specific and are determined automatically. This illustrates that the method and the associated web application can still be improved.

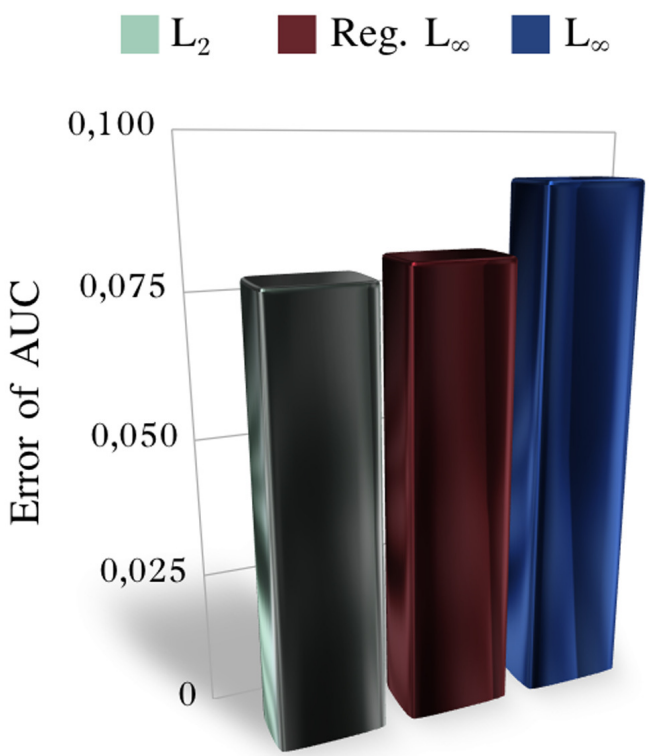

Figure 5.1: Comparison of three optimization algorithms. The error of AUC (one minus the AUC, or Area Above the Curve, AAC) is plotted on the y-axis, the lower the better. The three optimization algorithms are (i) L_∞, described and implemented in De Bie *et al.* that calculates sparse solutions for the kernel coefficients (grey, left bar), (ii) a regularized version of L_∞, in which a minimal boundary is set for the kernel coefficients (red, middle bar), and (iii) L_2, that can calculate non sparse solutions as in Yu *et al.*, (2010) (blue, right bar). L_2 and regularized L_∞ achieve similar performance (error of 0.0780 and 0.0806 respectively) while L_∞ performs significantly worse (0.0923, p-value < 0.001, paired t-test).

Chapter 6

Cross-species candidate gene prioritization with MerKator

6.1 Summary

The computational method presented in the previous chapter (5) is based on the use of kernel methods to perform candidate gene prioritization. It is shown to outperform our first strategy based on Order Statistics (chapter 3). The present chapter describes a cross-species prioritization strategy that is based on our previously described kernel based method augmented with a Noisy-Or model responsible of the cross-species data integration. This chapter also presents 'MerKator', a software that implements this strategy for five organisms, and discusses the computational challenges that this represents (*e.g.*, kernel computation, kernel centering, missing values).

Cross-species candidate gene prioritization with MerKator

Shi Yu[1], Léon-Charles Tranchevent[1], Sonia Leach[1], Bart De Moor[1], and Yves Moreau[1#]

December 20, 2010

[1] Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium.
[#] Corresponding author: Yves.Moreau@esat.kuleuven.be

Abstract

In biology, there is often the need to prioritize large list of candidate genes to further test only the most promising candidate genes with respect to a biological process of interest. In the recent years, many computational approaches have been developed to tackle this problem efficiently by merging multiple genomic data sources. We have previously described a gene prioritization method based on the use of kernel methods and proved that it outperforms our previous method based on order statistics. In the present paper, we report the extension of the method to support data integration over multiple related species and the development of a web based interface termed 'MerKator' that implements this strategy and proposes candidate gene prioritization for five species. Our cross-species approach has been benchmarked and cases studies demonstrate that human prioritizations can benefit from model organism data.

1 Introduction

In modern biology, the use of high-throughput technologies allows researchers and practicians to quickly and efficiently screen the genome in order to identify the genetic factors underlying a given disorder. However these techniques are often generating large lists of candidate genes among which only one or a few are actually associated to the disease of interest. Since the individual validation of all these candidate genes is often too costly and time consuming, only the most promising genes are experimentally assayed. In the past, the selection of the most promising genes relied on the expertise of the researcher, and its *a priori* opinion about the candidate genes. In the post-sequence era, the use of high-throughput technologies has generated a large amount of complex data, that is analyzed using *in silico* methods. In the last decade, several methods have been developed to tackle the gene prioritization problem (recently reviewed in [12]). Most of them combine genomic knowledge with pure experimental data to leverage the effect between reliability and novelty. For instance, POCUS, one of the earliest gene prioritization solution was developed by Turner *et al.*

in 2003 [14]. POCUS relies on Gene Ontology annotations, InterPro domains, and expression profiles to identify the genes potentially related to the biological function of interest. The predictions are made by matching the Gene Ontology annotations, InterPro domains and expression profiles of the candidate genes to the ones of the genes known to be involved in the biological function of interest. The POCUS system favors the candidate genes that exhibit similarities with the already known genes. Most of the proposed prioritization methods also rely on this 'guilt-by-association' concept.

Most of the existing prioritization approaches are restricted to integrating information in a single specie. However, people have recently started to collect evidence among multiple species to facilitate the prioritization of candidate genes. Chen *et al.* proposed ToppGene that performs prioritization for human based on human data (*e.g.*, functional annotations, proteins domains) as well as mouse data (*i.e.*, phenotypic data) [4]. Through an extensive validation, they showed the utility of mouse phenotypic data in human disease gene prioritization. Hutz *et al.* [5] have developed CANDID, an algorithm that combines cross-species conservation measures and other genomic data sources to rank candidate genes that are relevant to complex human diseases. Liu *et al.* have investigated the effect of adjusting gene prioritization results through cross-species comparison. They identified the ortholog pairs between *Drosophila melanogaster* and *Drosophila pseudoobscura* using BLASTP and used this cross-species information to adjust the rankings of the annotated candidate genes in *D. melanogaster*. They report that a candidate gene with a lower score in the main species (*D. melanogaster*) can efficiently be re-ranked higher if it exhibits a strong sequence similarity to ortholog genes [6]. According to their evaluation on 7777 loci of *D. melanogaster*, the cross-species model outperforms other single species models in sensitivity and specificity measures. Another related method is STRING developed by von Mering *et al.* [15]. STRING is a database that integrates multiple data sources from multiple species into a global network representation. STRING allows users to look at the interactions between genes from one specie using data from 630 organisms.

In this paper, we present MerKator, whose main feature is the cross-species prioritization through genomic data fusion over multiple data sources and multiple species. This software is developed on the Endeavour data sources [1, 13] and a kernel fusion novelty detection methodology [3]. Our approach is different from previous approaches since our cross-species integration scheme is not limited to a single data source nor to a single specie. At the contrary, MerKator can integrate 14 genomic data sources over 5 species. To our knowledge, MerKator is also the first candidate gene prioritization software powered by kernel methods. In this paper, we present and discuss the computational challenges inherent to such implementation. We also present a benchmark analysis, through leave-one-out cross-validation, that shows the efficiency of the cross-species approach.

2 Materials and Methods

2.1 Data sources

The goal of MerKator is to facilitate the understanding of human genetic disorders using genomic information across organisms. MerKator identifies the

Table 1: Genomic data sources adopted in Endeavour MerKator

data source	H.sapiens	M.musculus	R.norvegicus	D.melanogaster	C.elegans
Annotation GO	√	√	√	√	√
Annotation-Interpro	√	√	√	√	√
Annotation-EST	√	√			
Sequence-BLAST	√	√	√	√	√
Annotation-KEGG	√	√	√	√	√
Expression-Microarray	√	√	√	√	√
Annotation-Swissprot		√	√	√	√
Text	√				
Annotation-Phenotype				√	
Annotation-Insitu				√	
Motif	√				
Interaction-Bind	√				
Interaction-Biogrid	√			√	√
Interaction-Mint				√	√

homologs of *Homo sapiens* genes, considered as the main organism, in four reference organisms: *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. The identification is based on NCBI's HomoloGene [8, 16, 17], which provides the mapping of homologs among the genes of 18 completely sequenced eukaryotic genomes. For each gene in each organism, MerKator stores the homolog pair with the lowest ratio of amino acid differences (the *Stats-prot-change* field in HomoloGene database). MerKator incorporates 14 genomic data sources in multiple species for gene prioritization. The complete list of the data sources adopted in the current version is presented in Table 1.

2.2 Kernel workflow

MerKator uses a one class SVM [9, 11, 3] to obtain prioritization scores within a single organism. Then the prioritization scores obtained from multiple species are integrated using a Noisy-Or model. As mentioned, MerKator is a real bioinformatics software powered by kernel methods therefore many challenges are tackled in its design and implementation. Considering the efficiency of kernel methods implemented in real full-genomic scale application, MerKator separates the program into the offline process and the online process to improve its efficiency.

2.2.1 Approximation of kernel matrices using the incomplete Cholesky decomposition

The main computational burden is the kernel computation of various data sources in the full genomic scale, especially for the data that is represented in high dimensional space, such as Gene Ontology annotations, gene sequences, and text-mining among others. To tackle this difficulty, MerKator manages all the kernel matrices in an offline process using a Matlab-Java data exchange tool. In Matlab, the tool retrieves the genomic data from the databases and construct the kernel matrices. The kernel matrices of the full genomic data may be very large so it is not practical to handle them directly computationally. To solve this, we decompose all the kernel matrices with ICD (Incomplete Cholesky Decomposition), thus the dimensions of the decomposed kernel matrices are often

smaller than the original data. In MerKator, the precision of the ICD is set as 95% of the matrix norm, given by

$$\frac{||K - K'||_2}{||K||_2} \leq 0.05, \tag{1}$$

where K is the original kernel matrix, K' is the approximated kernel matrix as the inner product of the ICD matrix. In this way the computational burden of kernel calculation is significantly reduced as the computation of the inner product of the decomposed matrices. The Matlab-Java tool creates Java objects on the basis of decomposed kernel matrices in Matlab and stores them as serialized Java objects. The kernel computation, its decomposition and the Java object transformation are computationally intensive processes, and so they are all executed offline. For the online process, MerKator loads the decomposed kernel matrices from the serialized Java objects, reconstructs the kernel matrices and solve the 1-SVM MKL optimization problem to prioritize the genes as already described in De Bie *et al.* [3]. Then the prioritization results are displayed on the web interface. In contrast with the offline process, the online process is less computational demanding and the complexity is mainly determined by the d number of training genes ($O(d^3)$). In our implementation, the optimization solver is based on the Java API of MOSEK[2] that shows satisfactory performance (presented in Results and Discussion).

2.2.2 Kernel centering

In MerKator, when the prioritization task involves data set of the full genomic size, some trivial operations become quite inefficient. To control the number of false positive genes in 1-SVM, De Bie *et al.* suggest a strategy to center the kernel matrices that contain both the training genes and the test genes on the basis of the *iid* assumption. As mentioned in the work of Shawe-Taylor and Cristianini [10], the kernel centering operation expressed on the kernel matrix can be written as

$$\hat{K} = K - \frac{1}{l}\mathbf{1}\mathbf{1}^T K - \frac{1}{l}K\mathbf{1}\mathbf{1}^T + \frac{1}{l^2}(\mathbf{1}^T K \mathbf{1})\mathbf{1}\mathbf{1}^T, \tag{2}$$

where l is the dimension of K, $\mathbf{1}$ is the all 1s vector, T is the vector transpose. Unfortunately, when the task is to prioritize the entire genome, centering the full genome kernel matrices becomes inefficient. For MerKator, we use a strategy based on the split of the full genomic data into smaller subsets. Let us assume that the full genome data contains $\mathcal{N}$ genes, and is split into several subsets containing $\mathcal{M}$ genes. Instead of centering the kernel matrix sizes of $\mathcal{N} \times \mathcal{N}$, we center the kernel matrix of size $\mathcal{A} \times \mathcal{A}$, where $\mathcal{A}$ is the number of genes in the union of the $\mathcal{M}$ candidate genes with the training genes. Because $\mathcal{M}$ is smaller than $\mathcal{N}$, for each centered kernel matrix MerKator obtains the prioritization score of $\mathcal{M}$ candidate genes, so it need to iterate multiple times (denoted as k, which is the smallest integer larger than $\frac{\mathcal{N}}{\mathcal{M}}$) to calculate the scores of all the $\mathcal{N}$ candidate genes. According to the *iid* assumption, if $\mathcal{M}$ is large enough then centering the kernel matrix of size $\mathcal{A} \times \mathcal{A}$ is statistically equivalent to centering the kernel matrix of the full genome, thus the prioritization scores obtained from the k iterations can precisely approximate the values obtained when centering the full genome data. Therefore, we may compare the prioritization scores of

the $\mathcal{N}$ genes even if they are obtained from different centered matrices, and thus we can prioritize the full genome. All these assumptions come to one non-trivial question: how to select an appropriate $\mathcal{M}$ for an efficient trade off between the reliability of *iid* assumption and the computational efficiency? In MerKator, $\mathcal{M}$ is determined via experiments conducted on the text mining data source with a set of 20 human training genes. First, a prioritization model is built and the 22743 human genes are scored by centering the linear kernel matrix of the full genome. The obtained values are regarded as the true prioritization scores, denoted as f. We also calculate the overall computation time, denoted as t. To benchmark the effect of $\mathcal{M}$, we try 10 different values from 1000 to 10000. In each iteration, the 20 training genes are mixed with $\mathcal{M}$ randomly selected candidate genes and the prioritization scores of the candidate genes are computed by centering the small kernel matrix. In the next iteration, we select $\mathcal{M}$ new candidate genes until all the 22743 genes are prioritized. The prioritization scores obtained by centering this small kernel matrix are denoted as f', and the computation time is also compared. The difference (error) between the prioritization scores obtained in these two approaches represents how well the $\mathcal{M}$ candidate genes approximate the *iid* assumption of the full genome, and is given by

$$e = \frac{||f - f'||_2}{||f||_2}. \tag{3}$$

We use this difference to find the optimal $\mathcal{M}$. According to the benchmark result presented in Supplementary Table 2, large $\mathcal{M}$ values lead to small error but take much longer time for the program to center the kernel matrix. In MerKator, we set the $\mathcal{M}$ to 4000, which represents a balance between a low error ($e < 0.05$) and a fast computing time (16 times faster than centering the full genome).

2.2.3 Missing values

In bioinformatics applications, clinical and genomic datasets are often incomplete and contain missing values. This is also true for the genomic data sources that underly MerKator, for which a significant number of genes are missing. In MerKator, the missing gene profiles are represented as zeros in the kernel matrices mainly for computational convenience. However, zeros still contain strong information so that they may lead to imprecise prioritization scores. In MerKator, kernel matrices are linearly combined to create the global kernel that is used to derive the prioritization scores. In order to avoid relying on missing data for this calculation (and therefore to favor the well studied genes), we use a strategy illustrated in Supplementary Figure 1 to combine kernel matrices with missing values. This strategy is similar to what is done within Endeavour. For a given candidate gene, only the non-missing based scores are combined to calculate the overall score. The combined kernel matrix obtained by this strategy is still a valid positive semi-definite kernel and thus the obtained prioritization scores only rely on the non-missing information.

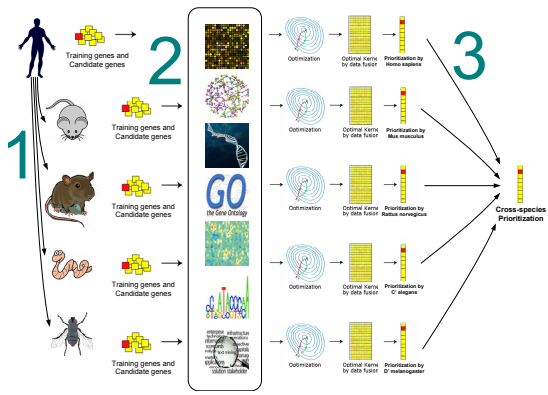

Figure 1: Conceptual overview of Endeavour MerKator software.

2.3 Cross-species integration of prioritization scores

MerKator first uses the one class SVM (1-SVM) algorithm to prioritize genes in a single specie, and then adopts a Noisy-Or model [7] to integrate prioritization scores from multiple species. The integration model is given as follows. We assume the scenario of cross-species prioritization as depicted in Figure 1. Similar to Endeavour, MerKator takes a machine learning approach by building a disease-specific model on a set of disease relevant genes, denoted as *training set*, then that model is used to rank the candidate genes, denoted as *candidate set*, according to their similarities to the model.

Some fundamental notions in cross-species gene prioritization are illustrated in Figure 2. Suppose in the main organism we specify N_1 number of human genes as $\{H_1, ..., H_{N_1}\}$ and MerKator obtains the corresponding training sets in reference species rat and mouse. The training set of rat contains N_2 genes as $\{R_1, ..., R_{N_2}\}$ and the training set of mouse has N_3 genes as $\{M_1, ..., M_{N_3}\}$. Note that MerKator always selects the homolog with the highest similarity ratio of sequence, so it is a *many-to-one* mapping thus N_2, N_3 are always smaller or equal to N_1. We define the homolog scores between the training sets of human and rat as $a_1, ..., a_{N2}$; Similarly, the homolog scores between human and mouse training sets are $b_1, ..., b_{N3}$. For the candidate set, each candidate gene of human is mapped to at most one rat gene and one mouse gene, where the homolog score is respectively denoted as c_0 and d_0. The homolog genes and the associated scores are all obtained from the NCBI HomoloGene database (release 63). To calculate the cross-species prioritization score, we introduce a set of utility parameters as follows.

We denote $h1$ and $h2$ as the parameters describing the quality of the homology, given by:

$$h1 = \min\{c_0, \text{median}(a_1, a_2, ..., a_{N2})\},$$
$$h2 = \min\{d_0, \text{median}(b_1, b_2, ..., b_{N3})\}.$$

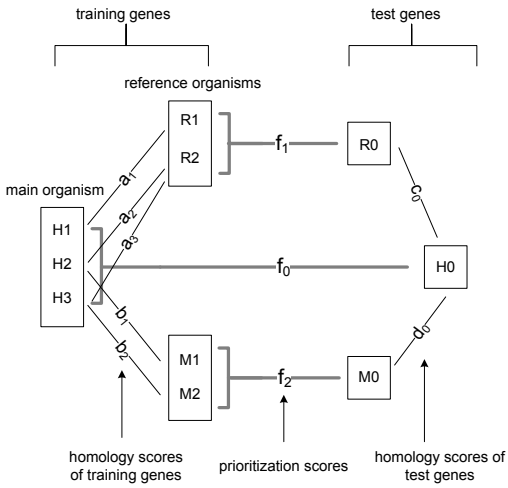

Figure 2: Integration of cross-species prioritization scores.

z_1 and z_2 are denoted as the parameters describing the ratio of the number of homologs in the reference organism with the number of genes in the main organism, given by

$$z_1 = \frac{N_2}{N_1}, \quad z_2 = \frac{N_3}{N_1}. \tag{4}$$

Next, we denote f_0 as the prioritization score of candidate gene H_0 ranked by the training set $\{H_1, ..., H_{N1}\}$; denote f_1 as the score of reference candidate gene R_0 prioritized by the reference training set $\{R_1, ..., R_{N2}\}$ and f_2 as the score of M_0 ranked by the set $\{M_1, ..., M_{N3}\}$. The raw prioritization scores obtained by 1-SVM are in the range of $[-1, +1]$, thus we scale them into $[0, +1]$. The adjustment coefficient adj is defined as:

$$adj = 1 - \prod_{\text{organism } i} (1 - h_i z_i f_i)\,. \tag{5}$$

The adj coefficient combines information from multiple species by the Noisy-Or model. A larger adj means there is strong evidence from the homologs that the candidate gene is relevant to the model. Considering the case one may want to eliminate the homolog bias, we further correct the adj parameter, denoted as adj^+, given by

$$adj^+ = \begin{cases} \text{median}(\{adj\}), & \text{if } j \text{ has no homolog} \\ 1 - \{\prod_{\text{organism } i} (1 - h_i z_i f_i)\}^{\frac{1}{k}} & \text{if } j \text{ has homolog(s).} \end{cases} \tag{6}$$

The first case of equation (6) means that when gene j has no homolog related, its adj^+ score equals to the median value of the set $\{adj\}$ that contains the adjustment values of the genes that have at least one homolog gene. In the second case, when there are k number of homologs mapped to gene j, we use the k-th exponential root removes the additional bias of the prioritization score caused by the multiple homologs (as shown in Supplementary Note 1).

This coefficient adj^+ is used to adjust f_0, the prioritization score of the main organism, and we have tried two different versions as follows:

$$\text{human non-special: } f_{\text{cross-species}} = 1 - (1 - f_0)(1 - adj^+), \tag{7}$$

and

$$\text{human special: } f_{\text{cross-species}} = 1 - \frac{(1 - f_0)(2 - adj^+)}{2}. \tag{8}$$

The *human non-special* version considers the human prioritization score as equivalent to the homology evidence and combines them again using the Noisy-Or function. In contrast, the *human special* version only adjusts the human prioritization score with the homology evidence by average. In the Noisy-Or integration, the cross-species score is boosted up if either the main species or the homology evidence shows good prioritization score. In the average score integration, the cross-species score compromises between the homology evidence and the main species score, and is only boosted up if both of them are good.

3 Software structure and interface

This section presents the interface of the software and explains how MerKator works. Using MerKator, a prioritization can be prepared in 4 steps (see Figure 3). In the first step, the user has to define the main organism, it will be the reference organism and will be used to input the training and candidate genes. In addition, the user can select other organisms to use, the corresponding species specific data sources will then be included further in the analysis. If no other organism is selected, the results are only based on the main organism data sources. In a second step, the training genes are inputed. Genes from the main organism can be inputed using various gene identifiers (*e.g.*, EnsEMBL, gene name, EntrezGene) or even pathway identifiers from KEGG or Gene Ontology. In addition, for human, an OMIM entry number can be inputed. Genes are loaded into the system using the 'Add' button. In the third step, the data sources to be used are selected by checking the corresponding boxes. By default, only the data sources of the main organism are displayed and the program is automatically selecting the corresponding data sources in the reference organisms when available. To have a full control on the data sources, the user must enter the advanced mode by clicking the dedicated button. Using the advanced mode, data sources from other organisms can be selected individually. In the fourth step, the candidate genes to prioritize are inputed. The user has two possibilities, either use the whole genome (in case the results are returned by e-mail) or input a subset of the genome (in case results are displayed in the web interface). For the latter, the method is similar to the second step, but genomic regions can also be inputed (*e.g.*, band q11 on chromosome 22, or region 100k - 900k on chromosome 19). The prioritization can be launched from this panel.

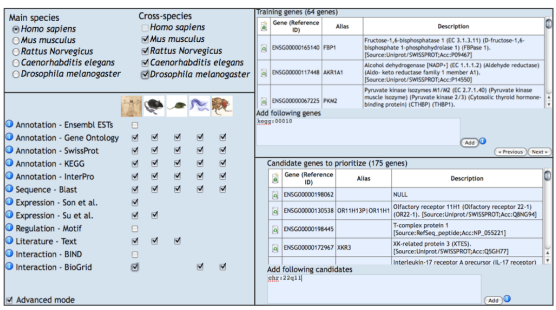

Figure 3: Typical MerKator workflow. This includes the species selection step (first step - top left), the input of the training genes (second step - top right), the selection of the data sources (third step - bottom left) and the selection of the candidate genes (fourth step - bottom right). Screenshots were taken from our online web server.

4 Results and discussion

The present paper introduces MerKator, a novel gene prioritization software based on kernel methods that can perform cross-species prioritization in five organisms (human, rat, mouse, fruit fly and worm). Compared to the previous approaches, our method differs by the number of organisms combined (current prioritization approaches focus either on mouse or fruit fly) as well as by the information that is combined (current prioritization approaches focus on conservation or expression data). The String approach of von Mering *et al.* is, to some respect, similar to our approach. The main differences with our method are, first, that String predicts novel interactions but does not perform prioritization, and, second, that String relies mostly on its text-mining component while we aim at integrating several genomic data sources (including but not restricted to text-mining).

To improve the efficiency of Merkator, we tackle the kernel computational challenges of full genomic data from multiple aspects. First, most of the computations was done offline and performed only once, restricting the case specific online computation to a strict minimum. Second, the prioritization of the full genome utilizes some approximation techniques such as incomplete Cholesky decomposition, kernel centering in the subsets of genome, and missing value processing to improve its feasibility and efficiency. Based on these efforts (details presented in Materials and Methods), Merkator is able to integrate all the adopted data sources from five species and prioritize the full human genome within 20 minutes.

We have developed a Noisy-Or based method to integrate the scores from multiple species into a global score. This Noisy-Or based method integrates the species specific scores by taking into account the strength of the homology between the corresponding species (see Figure 2). The use of a Noisy-Or based method is motivated by the fact that an excellent prioritization score obtained in one species should be enough to obtain an overall excellent score, which other measures such as the average would not allow. We have developed two scoring schemes termed *human special* and *human non-special*. The first one assumes that the source species (human in our case) is the main organism and that the other species are only used to adapt the score obtained in the main organism. The contribution of the other species is a half in total (the other half is the main organism score). The second solution is however relying on the hypothesis that all the species can contribute evenly, the main organism is not distinguished from the others. We have implemented and analyzed the two methods. In addition, we have implemented and tested two formula for the adjustment coefficient, adj and adj^+, to account for the differences in number of homolog genes. We have observed that the adj coefficient can introduce a bias towards the genes that have multiple homologs as compared to the genes that do not have any homologs. Either the homolog genes are still unknown or there is no homolog in any of the other species and therefore the gene is a human specific gene. In both cases, there is no rationale behind the bias and the gene should get the same chance to rank high than the other genes.

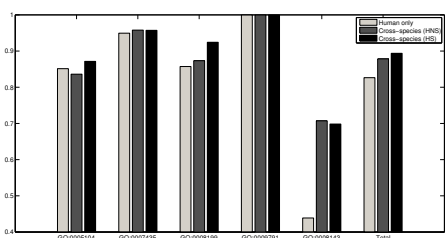

Figure 4: Benchmark results on five GO pathways using human only data sources (grey bars) and our cross-species data integration model (black bars).

As a proof of concept, we have benchmarked MerKator with five biological pathways (see Supplementary Table 3) using a leave-one-out cross-validation procedure. The four pathways were derived from Gene Ontology and contain a total of 37 genes. The validation was performed using all data sources except Gene Ontology and all five species. In this case, we have used the *human special* formula together with the adj^+ coefficient. Similar results are obtained with any other formula-coefficient combination (Supplementary Table 1). The Area Under the ROC Curve (AUC) is used as an indicator of the performance. We obtained a global AUC of 89.38% for the cross-species model, while the model based on human data alone obtains a smaller AUC of 82.64% (see Figure 4). For four out of the five pathways, the cross-species model performs better

than the human specific model although significance is not reached given the low number of genes per pathways (between 6 and 9). For the remaining pathway (GO:0008199), the two models achieve similar performance. This is because the human only performance is already too high to allow any significant improvement (AUC >99.9%). These results indicate that our cross-species model is conceptually valid and that reference organism genomic data can enhance the performance of human gene prioritization.

5 Conclusion

This paper presents MerKator, a software that combines cross-species information and multiple genomic data sources to prioritize candidate genes. The software is developed using the same databases adopted in Endeavour, but is equipped with a kernel fusion technique and a cross-species integration model. To embed kernel methods in a real and large scale bioinformatics application, we have tackled several computational challenges that are mentioned and discussed in this paper. Our approach may be concluded with the following three aspects:

- *Combining evidences from multiple species.* We proposed a Noisy-Or model to combine prioritization scores from multiple organisms. The issue of multiple species prioritization is complicated, which may involve many factors such as the size of training set, the selection of data sources, the number of relevant homologies, and so on. Considering so many factors, it is difficult to make statistical hypothesis, or estimate the data model for the final prioritization score. Thus our approach alternatively avoids the assumption about the data model of prioritization scores and calculates it using support vector machines. The integration methods are adjusted in the blackbox and the outputs are validated with benchmark data until satisfying performance is obtained.

- *User friendly interface.* Gene prioritization softwares are oriented to a specific group of computational biologists and medical researchers, therefore we designed an user friendly interface that is similar to Endeavour's web interface, and that does not require advanced mathematical skills to be used (configuration of the 1-SVM and the integration models are transparent to the end users). The results of full genomic cross-species prioritization are either directly returned or stored on the server and delivered to the end-user by e-mail messages depending on the number of candidate genes. When receiving the email notice, the users can either upload the prioritization results and display them in MerKator or download the results in XML format to extract the relevant information by themselves.

- *Near optimal solution.* The performance of kernel-based algorithms is strongly affected by the selection of hyper-parameters, such as the parameter of kernel function, or the regularization parameter. The optimal parameters should be selected by cross-validation, which may not be always feasible for a software oriented for biologists and medical researchers. Kernel fusion techniques allow developers to preselect the kernel parameters empirically. The overall performance does not rely on a single kernel

parameter, so even when the optimal parameter is not involved, the fusion procedure still can leverage among several near optimal parameters and provides a near optimal result. For real applications, the 1% difference of performance is not so critical to the end users. In most cases, a successful application prefers much the speed of solution than the very optimality of the parameter or the model.

Future work includes, but is not restricted to, the inclusion of more species and more data sources, the development of new modules to enhance even further the performance of our kernel based prioritization algorithm, the parallelization of computational methods to incorporate more data sources and more species, and the application to real biological problems, for instance through the integration of MerKator into research workflows.

6 Funding

This research is funded by the Research Council KUL (ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys, START 1), the Flemish Government: FWO (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 (Silicos, SBO-BioFrame, SBO-MoKa, TBM-IOTA3), FOD (Cancer plans). Additional funding agencies include the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011) and the EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED.

References

[1] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Léon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, Peter Carmeliet, and Yves Moreau. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544, May 2006. PMID: 16680138.

[2] E D Andersen and K D Andersen. The mosek interior point optimizer for linear programming: an implementation of the homogeneous algorithm. *High Perf. Optimization*, 2000.

[3] Tijl De Bie, Léon-Charles Tranchevent, Liesbeth M M van Oeffelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics (Oxford, England)*, 23(13):i125–132, July 2007. PMID: 17646288.

[4] Jing Chen, Huan Xu, Bruce J Aronow, and Anil G Jegga. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, 8:392, 2007. PMID: 17939863.

[5] Janna E Hutz, Aldi T Kraja, Howard L McLeod, and Michael A Province. CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genetic Epidemiology*, 32(8):779–790, December 2008. PMID: 18613097.

[6] Qian Liu, Koby Crammer, Fernando C N Pereira, and David S Roos. Reranking candidate gene models with cross-species comparison for improved gene prediction. *BMC Bioinformatics*, 9:433, 2008. PMID: 18854050.

[7] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[8] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael Dicuccio, Scott Federhen, Michael Feolo, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J Lipman, Zhiyong Lu, Thomas L Madden, Tom Madej, Donna R Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrachi, James Ostell, Anna Panchenko, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Stephen T Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A Tatusova, Lukas Wagner, Yanli Wang, W John Wilbur, Eugene Yaschenko, and Jian Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 38(Database issue):D5–16, January 2010. PMID: 19910364.

[9] B Schölkopf, JC Platt, J Shawe-Taylor, AJ Smola, and RC Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[10] J Shawe-Taylor and N Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[11] DMJ Tax and RPW Duin. Support vector domain description. *Pattern Recognition Letter*, 20:1191–1199, 1999.

[12] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau. A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 2010.

[13] Léon-Charles Tranchevent, Roland Barriot, Shi Yu, Steven Van Vooren, Peter Van Loo, Bert Coessens, Bart De Moor, Stein Aerts, and Yves Moreau. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Research*, 36(Web Server issue):W377–384, July 2008. PMID: 18508807.

[14] Frances S Turner, Daniel R Clutterbuck, and Colin A M Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biology*, 4(11):R75, 2003. PMID: 14611661.

[15] Christian von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue):D433–437, January 2005. PMID: 15608232.

[16] D L Wheeler, D M Church, A E Lash, D D Leipe, T L Madden, J U Pontius, G D Schuler, L M Schriml, T A Tatusova, L Wagner, and B A Rapp.

Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 29(1):11–16, January 2001. PMID: 11125038.

[17] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David L Kenton, Oleg Khovayko, David J Lipman, Thomas L Madden, Donna R Maglott, James Ostell, Kim D Pruitt, Gregory D Schuler, Lynn M Schriml, Edwin Sequeira, Stephen T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tugba O Suzek, Roman Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 34(Database issue):D173–180, January 2006. PMID: 16381840.

6.2 Contribution of the PhD candidate

The PhD candidate has gathered the genomic data sources used to build the kernels. He has developed the web interface. He has also performed the benchmark analysis. He has participated in the writing of the paper.

6.3 Discussion

6.3.1 Network based strategy

An alternative to kernel based methods might reside in network based methods. Similarly to kernels, almost every genomic data source can be represented as a network in which the nodes represents the genes or the proteins and the edges the links that exist between them. This representation of the genomic data as networks permits the use of network based methods to prioritize candidate genes. Several of these methods are very similar to the kernels based methods since there exist similarities between the kernel and the network representations.

In a preliminary study performed in collaboration with Daniela Nitsch, we used a network representation to prioritize candidate genes using expression data for training and illustrate on 4 monogenic disorders [169]. In this study, we assessed a candidate gene by considering the differential expression of its neighborhood in a gene network under the assumption that strong candidates will tend to be surrounded by differentially expressed neighbors. The gene network is built by considering several genomic data sources from several organisms [246] and by applying a diffusion (laplacian exponential diffusion) to include indirect links into the network.

One of the main difference with existing methods is the use of expression data to represent the disease under study. The existing methods are either using known disease genes or keywords that describe the disease to train their models, this novel strategy introduces a novel possibility through the use of an expression data set that contains the differential expression level of the genes in a disease sample (as compared to a reference sample).

Chapter 7

Large-scale benchmark of Endeavour using MetaCore maps

7.1 Summary

The benchmark of our algorithms is as important as their development, since benchmarking is a first step towards experimental validation. Originally, Endeavour was benchmarked by leave-one-out cross-validation on 32 gene sets corresponding to 3 bio-molecular pathways and 29 genetic diseases, representing 695 prioritizations in total. Although very useful to estimate the performance on real biological question, this only represents a small fraction of the scientific knowledge of genetic diseases and bio-molecular pathways. In this chapter, a larger benchmark using 1276 pathway maps and disease marker sets from MetaCore™, totalizing 22343 prioritizations is reported. This benchmark has been realized in the context of a collaboration with one of the major international pharmaceutical company, Novartis Pharma AG. In particular, the prioritizations were mainly run at Novartis without fine tuning the prioritization system. Results show that the hypothesis we derive from our previous small scale benchmark also stands for larger benchmarks.

Bioinformatics Advance Access published June 10, 2010

Large-scale benchmark of Endeavour using MetaCore maps

Sven Schuierer[1,*], Léon-Charles Tranchevent[2,3,*], Uwe Dengler[1] and Yves Moreau[2,3]

[1]Novartis Pharma AG, Postfach, CH-4002 Basel, Switzerland.

[2]Departement of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium.

[3]SymBioSys Center for Computational Systems Biology, Katholieke Universiteit Leuven, Leuven, Belgium.

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Summary: Endeavour is a tool that detects the most promising genes within large lists of candidates with respect to a biological process of interest and by combining several genomic data sources. We have benchmarked Endeavour using 450 pathway maps and 826 disease marker sets from MetaCore™ of GeneGo, Inc containing a total of 9,911 and 12,432 genes respectively. We obtained an AUC of 0.97 for pathway and of 0.91 for disease gene sets. These results indicate that Endeavour can be used to efficiently prioritize candidate genes for pathways and diseases.

Availability: Endeavour is available at http://www.esat.kuleuven.be/endeavour

Contact: Sven.Schuierer@novartis.com or Leon-Charles.Tranchevent@esat.kuleuven.be

1 INTRODUCTION

Identifying disease causing genes is a key challenge in human genetics. In the process of identifying such disease genes, researchers are often confronted with large lists of candidate genes among which only one or a few are actually causal. The validation of each candidate is often too costly and time consuming, so that only a few candidates are further experimentally validated. A related problem arises when trying to identify new members of a biological pathway. The selection of a small subset of optimal candidates for validation is called gene prioritization. Since going manually through all possible sources of information is a slow and tedious process, several bioinformatics methods have been developed to tackle this problem (Zhu and Zhao, 2007; Oti and Brunner, 2007). We previously developed Endeavour (Aerts *et al.*, 2006; Tranchevent *et al.*, 2008) whose key feature is that it uses multiple genomic data sources (e.g., sequence, expression, literature, annotation) to estimate how promising a candidate gene is by measuring its similarity with a set of training genes. The training genes are genes which are already known to play a role in the biological process under study. The underlying assumption is that the most promising candidate genes are the ones that exhibit many similarities with the training genes. A schematic view of the algorithm is shown on Fig. 1. Originally, Endeavour was benchmarked by leave-one-out cross-validations on 32 gene sets corresponding to 3 bio-molecular pathways and 29 genetic diseases, representing around 700 prioritizations in total (Aerts *et al.*, 2006). In the current study, we briefly report on the largest benchmark to date for a gene prioritization method using 1276 pathways and diseases from MetaCore and prioritizing a total of 22,343 genes.

*To whom correspondence should be addressed.

2 METHODS

We used the MetaCore™ Pathway Maps and Disease Marker Sets as provided by GeneGo, Inc in October 2008. This resulted in 450 pathway maps containing a total of 9,991 genes, and 826 disease marker sets containing a total of 12,432 genes (see also Supplementary Material). In addition, the OMIM and Gene Ontology based benchmarks were built as described in Aerts *et al.* (2006), see also Supplementary Material. The Endeavour prioritization platform was accessed remotely using a secured connection from a command line interface allowing the automatic processing of thousands of prioritizations.

3 RESULTS

The cross-validation procedure measures the ability of the program to capture the information of the known genes and to correctly use this information to prioritize the left-out gene. To assess the ability of Endeavour to capture the information of known pathway and disease-related gene sets, we used the pathways maps and disease marker sets of MetaCore™ from GeneGo, Inc. Since the gene sets in MetaCore are manually curated, we consider them as a reliable representation of the current knowledge of the functional contexts in which the genes are active. We have benchmarked Endeavour using 450 pathway maps and 826 disease marker sets containing a total of 9,991 pathway members and 12,432 disease genes respectively. In addition, we have also benchmarked 29 OMIM diseases and 37 Gene Ontology pathways that contain respectively 620 and 1216 genes. For each prioritization run, the position of the left-out gene among 99 randomly chosen candidates is recorded gene (see also Supplementary Material). We use the area under the Receiver Operating Characteristic (ROC) curves (AUC) as a measure of the performance. We obtained an AUC of 0.97 for the MetaCore pathways. Moreover, 64% of the prioritizations have the left-out gene being ranked in the first position. The AUC value obtained for the MetaCore disease marker sets is 0.91 and 33% of the prioritizations have the left-out gene being ranked in the first position (see also Fig. 2). The AUC values obtained for the Gene Ontology

pathways and OMIM diseases are respectively 0.93 and 0.94. Altogether, the results indicate that Endeavour efficiently prioritizes candidate genes for both pathways and diseases. As observed and discussed in our previous work (Aerts *et al.*, 2006), the performance of gene prioritization is higher for pathways than for disease marker sets because data sources such as Gene Ontology contain pathway specific information and because diseases often implicate a complex set of cascades making their profiling more challenging. Furthermore, the performance is higher for OMIM diseases than for MetaCore diseases because the MetaCore sets include markers derived from gene expression studies whereas the OMIM sets only rely on known causative genes. Such markers are indirectly associated to the disease and it is therefore harder to prioritize them. Assessing the performance of a novel type of bioinformatics tool, such as gene prioritization methods, is of crucial importance. Our large-scale benchmark demonstrates the effectiveness of Endeavour. It should be noted that the evaluation was carried out at Novartis by S. Schuierer and U. Dengler independently of the core Endeavour team. In particular, the Endeavour platform was used as is and no parameter fine tuning was performed (i.e., all available data sources were used, see also Supplementary Material). We are aware of the many pitfalls of benchmarking gene prioritization and function prediction methods (Myers *et al.*, 2006), so that the performance observed in cross-validation studies is likely to be higher than that observed in prospective studies. We have recently conducted such a prospective validation in Drosophila (Aerts *et al.*, 2009), which also confirmed further the effectiveness of our strategy.

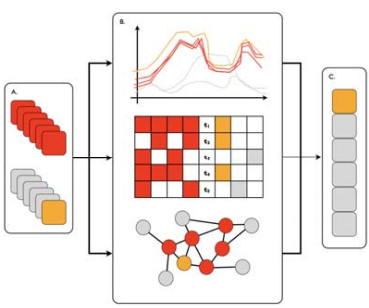

Fig. 1. The Endeavour algorithm. A. The inputs are, on the one hand, the training genes (on top - in red), known to be involved in the process of interest, and, on the other hand, the candidate genes to prioritize (at the bottom – in grey and orange). B. Data are collected for these genes: e.g., expression profiles, functional annotations, and protein-protein interactions. C. Candidate genes are prioritized, i.e., ranked according to their similarities to the training genes. For example, the gene in orange is the most promising candidate (i.e., it ranks in first position) because (i) its expression profile is similar to the red ones, (ii) it also shares several functional annotations, and (iii) it is interacting with several training proteins.

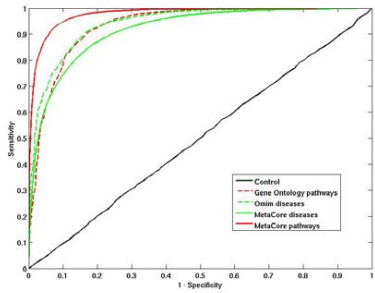

Fig. 2. Results of the large-scale validation of Endeavour on the 450 pathways and 826 disease marker sets from MetaCore. The disease receiver operating characteristic (ROC) curve, in green, results in an AUC of 91.65% and the pathway ROC, in red, indicates an even better performance with an AUC of 97.72%. The dotted curves represent the performance for the OMIM diseases (dotted green – 94,12%) and the GO pathways (dotted red – 93,37%). The black curve serves as a control (49.86%). The optimal control experiment would consist of shuffled gene sets but randomly selected gene sets were used as an approximation. AUCs for diseases and pathways are significantly larger than the control AUC (Wilcoxon rank sum < 1e-6).

ACKNOWLEDGEMENTS

We want to thank July Bryant and Yuri Nikolsky of GeneGo, Inc for many helpful discussions. We also want to thank Stefan Grzybek for his support during the project, and the anonymous reviewers for their thorough review of our manuscript.

Funding: This work was supported by the 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

Aerts,S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, 24: 537-544.

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

Myers,C.L. *et al.* (2006) Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7: 187.

Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, 71: 1-11.

Tranchevent,L.C. *et al.* (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, 36: W377-W384.

Zhu,M. and Zhao,S. (2007) Candidate gene identification approach: progress and challenges. *Int. J. Biol. Sci.*, 3: 420-427.

7.2 Contribution of the PhD candidate

The PhD candidate has developed a dedicated Java based client for Novartis with a secure connection between the local server and the Novartis client. He provided conceptual and technical support during the remote evaluation of the tool. The PhD candidate also performed locally three additional benchmarks (*i.e.*, control, Gene Ontology, and OMIM). He has also written the paper.

7.3 Discussion

To complete the benchmark presented in this chapter, 3 additional large scale benchmark sets have been produced using the Genetic Association Database (GAD), Kegg and Ingenuity®.

In total, this represents an addition of 94 Ingenuity® pathways, 147 Kegg pathways and 142 GAD diseases, for a total of 11777 genes. The ROC curves are presented in figure 7.1, indicating that our approach is able to efficiently validate these sets. We can observe that the pathway gene sets result in a better AUC (97,63% for Ingenuity® and 97,78% for Kegg) than the disease gene sets (92,58% for GAD), which is also true for the OMIM, GO and MetaCore™ gene sets. Of interest, the pathway validations for Kegg, Ingenuity®, MetaCore™ not only give approximately the same AUC but also very similar curves as can be appreciated from figure 7.1. This is most likely due to the overlap in the pathway gene sets that are coming from the distinct databases. This can also be observed for disease gene sets although the effect is not as strong as for pathways, the same explanation is valid since knowledge bases such OMIM and MetaCore™ rely on the same underlying scientific literature. In this case, more precisely OMIM is probably completely included in MetaCore™.

One of the weakest point of the LOOCV is the selection of the test genes, we are usually performing a random selection of 99 genes that constitutes the test set together with the left-out gene. Selecting a reduced set of 99 genes might induce a bias since they are unlikely representative of the whole genome, but repeating this procedure many times is thought to average things out. However, the small number of repetitions performed in our small scale benchmark might not be enough. To investigate this, we compare the results of the classical 99 random gene benchmark and the results of whole genome benchmark. Results are shown in figure 7.2 and indicate that there is no statistical difference for any of the data sources considered, nor for the overall performance. This means that for our prioritization experiments, taking 99 random genes to estimate the background is correct even with a small number of draws (*e.g.*, 620 for the OMIM based disease benchmark).

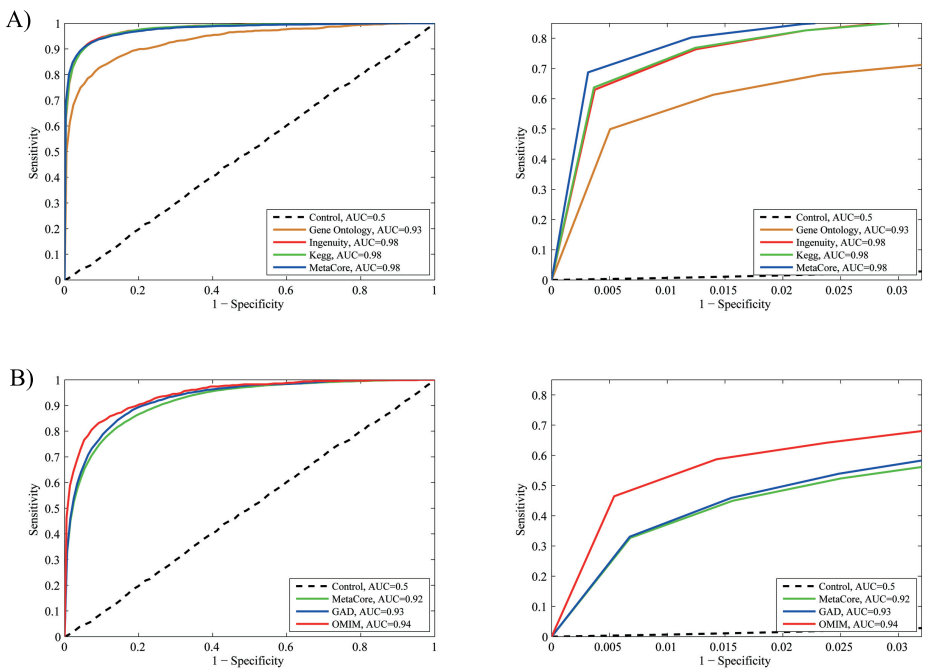

Figure 7.1: Results of a large scale benchmark analysis. Results are presented as ROC curves. (A) On the left, complete ROC curve for the pathway validation using Gene Ontology (GO, 93.37%), Kegg (97.78%), Ingenuity® (97.63%), and MetaCore™ (97.72%). On the right, zoom for the top 3% ranks to distinguish the overlapping curves. (B) On the left, complete ROC curve for the disease validation using OMIM (94.12%), the Genetic Association Database (GAD, 92.58%), and MetaCore™(91.65%). On the right, zoom for the top 3% ranks to distinguish the overlapping curves. Notice the control curves in dotted black obtained via the validation of randomly built gene sets. All the pathway and disease AUCs are significantly larger than the control AUC (Wilcoxon rank-sum < 0.001)

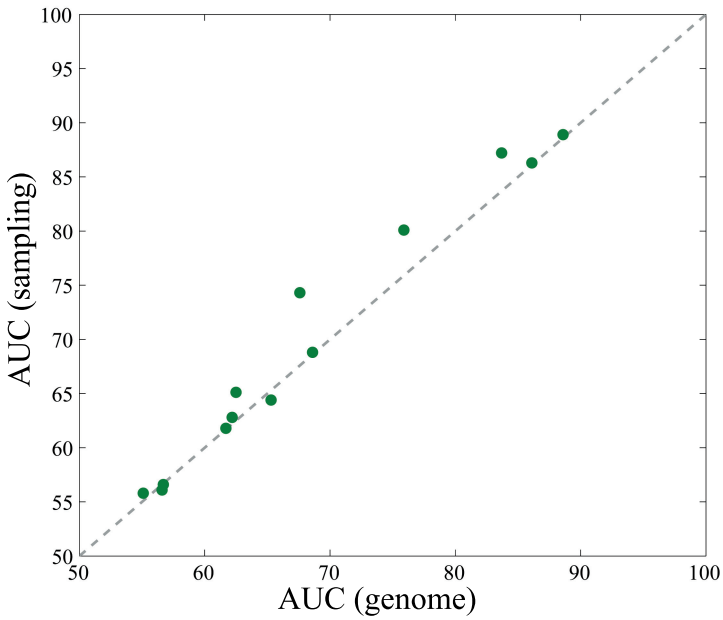

Figure 7.2: Results of a comparison between leave-one-out cross-validations performed on the whole genome and performed by sampling 99 candidate genes from the genome in each run. The AUC is reported for the OMIM based benchmark, for all models (*i.e.*, data sources), and for both setups (genome wide on the x-axis, with sampling on the y-axis). The results show that 99 randomly selected candidate genes can be used as a fast and relatively accurate estimate.of the genome.

Chapter 8

Integrating Computational Biology and Forward Genetics in Drosophila

8.1 Summary

The previous chapter (chapter 7) describes the benchmark of our candidate gene prioritization method. Benchmarking is an effective way to estimate the real performance, however a real experimental validation, *i.e.*, applying the method on real biological problems, is always more elegant. This chapter describes the development of a fly specific version of Endeavour termed 'Endeavour-HighFly', and its experimental validation through a computationally supported genetic screen performed in collaboration with the laboratory of Prof. Bassem Hassan.

Beside rat, mouse, and worm, fruit fly (*Drosophila melanogaster*) is a species that is often used to get more insight into developmental pathways [171, 18, 159, 25, 154]. The development of a fruit fly specific version includes the integration of 12 additional data sources including 2 large expression datasets. The fruit fly version is made available for both the Java server and the web server. Its is benchmarked through cross-validation on canonical signaling pathways indicating its validity for fly prioritization. Its was further experimentally validated through the development of a computationally supported genetic screen.

OPEN ACCESS Freely available online

PLoS GENETICS

Integrating Computational Biology and Forward Genetics in *Drosophila*

Stein Aerts[1,2☯], Sven Vilain[1,2,3☯], Shu Hu[1,2,3☯], Leon-Charles Tranchevent[4☯], Roland Barriot[4¤], Jiekun Yan[1,2], Yves Moreau[4], Bassem A. Hassan[1,2,3]*, Xiao-Jiang Quan[1,2☯]

1 Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, Vlaams Instituut voor Biotechnologie, Leuven, Belgium, **2** Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine, Leuven, Belgium, **3** Doctoral Program in Molecular and Developmental Genetics, Katholieke Universiteit Leuven School of Medicine, Leuven, Belgium, **4** Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium

Abstract

Genetic screens are powerful methods for the discovery of gene–phenotype associations. However, a systems biology approach to genetics must leverage the massive amount of "omics" data to enhance the power and speed of functional gene discovery *in vivo*. Thus far, few computational methods for gene function prediction have been rigorously tested for their performance on a genome-wide scale *in vivo*. In this work, we demonstrate that integrating genome-wide computational gene prioritization with large-scale genetic screening is a powerful tool for functional gene discovery. To discover genes involved in neural development in *Drosophila*, we extend our strategy for the prioritization of human candidate disease genes to functional prioritization in *Drosophila*. We then integrate this prioritization strategy with a large-scale genetic screen for interactors of the proneural transcription factor Atonal using genomic deficiencies and mutant and RNAi collections. Using the prioritized genes validated in our genetic screen, we describe a novel genetic interaction network for Atonal. Lastly, we prioritize the whole *Drosophila* genome and identify candidate gene associations for ten receptor-signaling pathways. This novel database of prioritized pathway candidates, as well as a web application for functional prioritization in *Drosophila*, called ENDEAVOUR-HIGHFLY, and the Atonal network, are publicly available resources. A systems genetics approach that combines the power of computational predictions with *in vivo* genetic screens strongly enhances the process of gene function and gene–gene association discovery.

Citation: Aerts S, Vilain S, Hu S, Tranchevent L-C, Barriot R, et al. (2009) Integrating Computational Biology and Forward Genetics in *Drosophila*. PLoS Genet 5(1): e1000351. doi:10.1371/journal.pgen.1000351

Editor: Stuart K. Kim, Stanford University Medical Center, United States of America

Received September 10, 2008; **Accepted** December 19, 2008; **Published** January 23, 2009

Funding: This work was supported by an FWO postdoctoral fellowship to SA, VIB, Impuls, CREA and GOA grants from K.U. Leuven, and FWO grant number G.0542.08N, to BAH and the following grants to YM: GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, FWO (G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, ICCoS, ANMMM, MLDM), IWT (GBOU-McKnow-E, GBOUANA, TAD-BioScope-IT, Silicos, SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBMIOTA3), IUAP P6/25, ERNSI, FP6-NoE Biopattern, FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

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

* E-mail: Bassem.Hassan@med.kuleuven.be

¤ Current address: Laboratoire de Microbiologie et Génétique Moléculaires, Université de Toulouse, Toulouse, France

☯ These authors contributed equally to this work.

Introduction

The demand by systems biology for *bona fide*, *in vivo* validated, biochemical interaction data and high quality functional annotations is much higher than the supply that geneticists are able to provide, principally because genetic approaches mainly focus on generating data on a gene-by-gene basis. On the other hand, computational predictions of gene function alone remain far from being accurate enough to be considered high-quality biological data. Integrated solutions, that combine the advantages of several approaches, should in theory provide both fast and physiologically relevant genetic data, while simultaneously increasing our understanding of biological processes. Genetic interactions in model organisms constitute a potentially invaluable source of *in vivo* interaction data for systems biology provided that throughput and speed can be increased. Currently, the number of known genetic interactions remains much smaller than the number of annotated physical interactions. For example, the BioGRID [1] database currently contains approximately 53,000 genetic interactions compared to almost 100,000 physical interactions.

Clearly, the power of genetic approaches is that they produce - by definition - data that is directly relevant in a living system. Genetic screens, either for specific phenotypes or for modifiers of gene function, are thus a valuable source of large-scale interaction data. However, the main disadvantage of large-scale genetic screens is that they are costly, labor intensive, and time consuming. Turning *in vivo* genetic screens into a staple of systems biology by making them easier and faster without compromising their accuracy would therefore represent a major advance.

In the bioinformatics community, process- or disease-related genes are, as of recently, being computationally predicted by taking advantage of the large amount of available sequence, function, annotation, and interaction data [2–13]. However to our knowledge, none of these methods have been used in combination with large-scale genetic experiments. Therefore, it remains unclear to what extent genome-wide, or even large-scale, computational predictions of gene-gene or gene-pathway associations, are biologically meaningful. Carrying out such screens on a large scale is difficult in human or mouse genetics, but the availability of genetic tools in *Drosophila melanogaster* together with collections of

Author Summary

Genome sequencing and annotation, combined with large-scale molecular experiments to query gene expression and molecular interactions, collectively known as Systems Biology, have resulted in an enormous wealth in biological databases. Yet, it remains a daunting task to use these data to decipher the rules that govern biological systems. One of the most trusted approaches in biology is genetic analysis because of its emphasis on gene function in living organisms. Genetics, however, proceeds slowly and unravels small-scale interactions. Turning genetics into an effective tool of Systems Biology requires harnessing the large-scale molecular data for the design and execution of genetic screens. In this work, we test the idea of exploiting a computational approach known as gene prioritization to pre-rank genes for the likelihood of their involvement in a process of interest. By carrying out a gene prioritization–supported genetic screen, we greatly enhance the speed and output of *in vivo* genetic screens without compromising their sensitivity. These results mean that future genetic screens can be custom-catered for any process of interest and carried out with a speed and efficiency that is comparable to other large-scale molecular experiments. We refer to this combined approach as Systems Genetics.

deficiency lines, mutants, and insertion lines, makes it an ideal model organism to investigate the concept of integrating genetic screens with gene prioritization methods.

Here, we integrate genetics and computational biology to identify genetic interactions underlying neural development in the *Drosophila* Peripheral Nervous System (PNS), a well-established model for neurogenesis. Proneural genes encoding proteins of the basic-helix-loop-helix (bHLH) super-family of transcription factors are essential for the initiation of neuronal lineage development in all species [14–18]. They act by forming heterodimers with the widely expressed bHLH E-proteins to bind a DNA motif called the E-box [19] and regulate the transcription of target genes. The highly conserved members of the Atonal (Ato) family are one example of proneural genes whose activity is required for the development of multiple lineages in vertebrates and invertebrates [14,20–22]. Despite a solid understanding of when and where *ato*-like genes are required in the *Drosophila* PNS and how they interact with Notch signaling to select neural precursor cells (NPCs), the mechanisms that mediate their activity within NPCs and their specificity in inducing neuronal differentiation remain largely obscure.

To identify genes involved in *ato* mediated neural development we propose a strategy for functional gene prioritization in *Drosophila* called ENDEAVOUR-HIGHFLY that uses the same data fusion method and user interface as the human gene prioritization method ENDEAVOUR [3,23]. We identify 18 genes that interact with *ato* in two different contexts, including 2 previously uncharacterized genes, and use them to predict a core Ato interaction network. Furthermore, to broaden our strategy to other developmental processes, we prioritize the entire *Drosophila* genome for each of ten canonical biological pathways and generate a freely available database of candidate members or interactors for each pathway.

Results

Identifying Modifier Loci of *atonal*-Induced Neurogenesis in the *Drosophila* PNS

Three amino-acids within the basic domain of the first helix have been shown to mediate the specificity of *ato* function [24], and the same motif enables specific transcriptional activation of the nicotinic acetylcholine receptor beta-3 subunit by the *ato* orthologue Ath5 [25]. Substituting the same amino acids in the Ato-related mouse proneural protein Neurogenin 1 (Ngn1) for Ato group-specific residues (Ngn^{bAto}) allows Ngn1 to induce neurogenesis in *Drosophila*. This induction mimics that caused by Ato itself and depends on the fly E-protein Daughterless (Da) and the proneural co-factor Senseless (Sens). Also, like endogenous proneural activity, it is antagonized by the Notch signaling pathway. Expression of the "Atonalized" form of mouse Ngn1, Ngn^{bATO} (Figure 1A) under the control of *dpp-Gal4* induces an average of ~30 ectopic sensory bristles on the adult wing vein ($n = 30$; Figure 1B, C). This is in contrast to an average of only ~7 bristles induced by Ngn1 itself ($n = 26$; $p<0.001$), but is similar to the number induced by Ato ($n = 26$, n.s.; Figure 1C). However, unlike Ato, Ngn^{bATO} induces significantly less lethality and many fewer wing deformities making it much easier to use in a large scale, quantitative genetic screen. In addition, just like for Ato, removal of one copy of *sens* reduces the number of Ngn^{bATO}-induced bristles by 55.6% (Figure 1C). In order to bring the screen to a dosage critical value, a heterozygous *sens* mutant was introduced into the background of *UAS::*Ngn^{bATO}*; dpp-Gal4*. The number of ectopic bristles with this system provides a sensitized and quantitative read out in which to screen for modifiers of Ato function.

To test the feasibility of isolating dominant modifiers of the number of ectopic bristles, we crossed *UAS::*Ngn^{bato}*/Cyo;sens,dpp-Gal4/TM6c*, flies to *da* or *Notch* mutant flies. We find that removal of a single copy of *da* almost completely suppressed Ngn^{bATO} induced bristle formation (average of 0.7 ± 0.9 bristles; $n = 27$, $p<0.001$), while removal of one copy of Notch strongly enhanced the phenotype (average of 43.5 ± 4.1 bristles; $n = 23$, $p = 0.002$; Figure 1C). All together, these data suggest that the assay is both robust and sensitive and should enable the identification of specific quantitative modifiers involved in *ato*-dependent neurogenesis in the *Drosophila* PNS.

Following this strategy, a deficiency screen of the second and the third chromosomes for modifiers of Ngn^{bato} misexpression was performed. The deficiency kit is a collection of fly stocks that each carries a deficiency, or deletion, chromosome uncovering multiple genes. The different deficiencies encompass most of the chromosome and deficiency screening is an established and rapid assay to identify chromosomal regions with enhancer and suppressor loci for a given phenotype or pathway [26]. To identify chromosomal loci that influence *ato*-induced neural development, 180 deficiency fly lines were crossed to *UAS::*Ngn^{bato}*/Cyo;sens,dpp-Gal4/TM6c*, flies. Loci were considered positive if they altered the number of ectopic bristles on the adult wing vein by more than 30% compared to the number of bristles induced in sibling control flies, as well as in wild type Canton S flies, and if the change in bristle number was strongly statistically significant ($p<0.01$). Following these stringent criteria, 17 positive regions on chromosome 2 and 14 positive regions on chromosome 3 were identified. Since induction of ectopic bristles is a common property of all proneural genes, the identified loci might be involved in both *achaete-scute* and *ato* dependent neurogenesis. In order to identify Ato-specific loci, the individual candidate deletion stocks were tested with flies expressing *UAS::ato*, *UAS::ngn1*, and *UAS::sc*, respectively, under the control of *dpp-Gal4*. The loci which modified Ato misexpression, but not that of Sc or Ngn1 were considered to be Ato-specific loci. Of the 31 loci identified in the primary screen, only one failed to interact with any of the genes in the secondary screen. We find that 15 of the 31 loci interact with both *ato* and at least one other proneural gene, while 2 loci interact only with *ngn1* and 1 locus interacts only with *sens* (data not shown). The remaining 12 loci (6

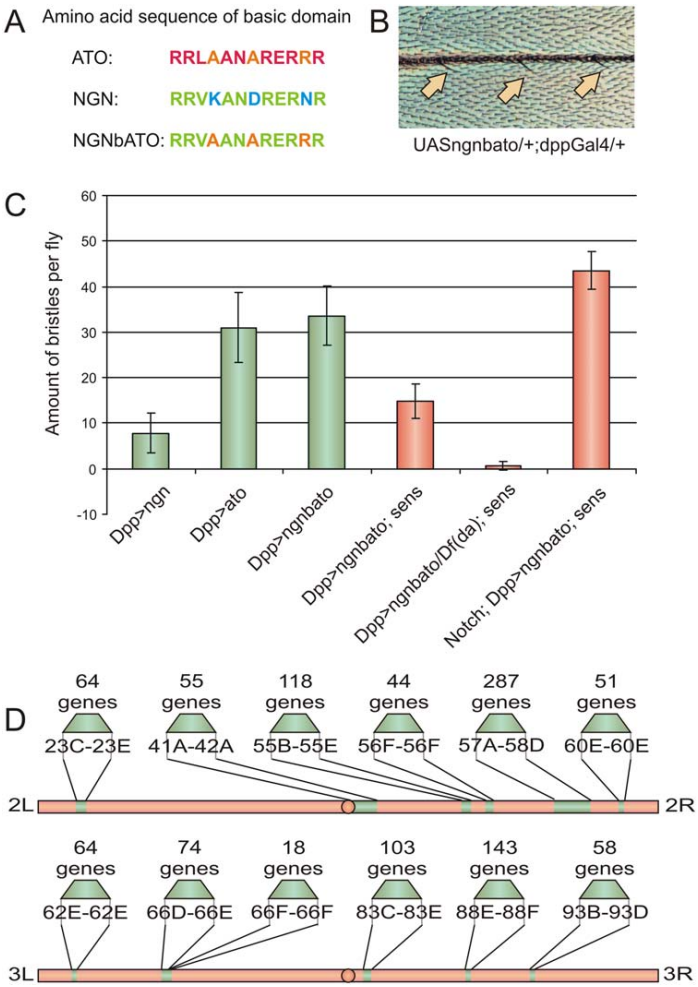

Figure 1. Overexpression of Ngn^{bato} in the *sens* mutant background provides a robust and sensitive phenotype for screening of *ato* dependent enhancers and suppressors. (A) Amino Acid sequence of the basic domain of Ngn (green) and Ato (red). The functionally critical amino acids are shown in separate colors. (B) Bristle phenotype on the third wing vein induced by Ngn^{bato} driven by dpp-Gal4. (C) Quantitative assay of ectopic bristle formation induced by Ngn^{bato} in wild type, *sens*/+, *da*/+;*sens*/+ and *N*/+; *sens*/+ backgrounds. Ato and Ngn are shown as positive and negative controls, respectively. Removing one copy of senseless reduces the amount of ectopic bristles. Removing one copy of *da* in a *sens*/+ background results in a suppression of the phenotype, whereas removing one copy of *N* results in an enhancement of the phenotype. (D) Cytological position of the deficiency regions and amount of genes found within each atonal positive deficiency region on chromosome 2 and chromosome 3.
doi:10.1371/journal.pgen.1000351.g001

on chromosome 2 and 6 on chromosome 3) interact specifically with *ato*. Examining the breakpoints of the overlapping deletions uncovering these 12 loci shows that they harbor 1056 annotated genes (Figure 1D and Table S1). Each of these loci is expected to harbor one or more *ato*-interacting genes.

The identification of the individual modifier genes from these regions is similar to the problem in human genetics where for a given human phenotype and its underlying chromosomal locus, identified by cytogenetic studies or linkage mapping for example, the individual disease-causing gene(s) need(s) to be identified.

Besides directly providing interaction candidates, the twelve positive regions resulting from the deficiency screen provide an excellent opportunity to test the principle of gene prioritization on a large scale and in an unbiased setup. First we present a redesign of an existing gene prioritization approach that is specifically tuned towards the *Drosophila* genome, and then we use it to select the most promising candidates from the 1056 genes within the twelve positive regions.

ENDEAVOUR-HIGHFLY: A Tool to Prioritize *Drosophila* Genes through Genomic Data Fusion

To prioritize *Drosophila* genes we upgraded the existing ENDEAVOUR tool for gene prioritization [3,23] by including *Drosophila* data sources (Table 1 and Materials and Methods) and we name this version ENDEAVOUR-HIGHFLY, or HIGHFLY for short. To test the performance of each individual *Drosophila* data source we carried out leave-one-out cross-validations (LOOCV; see Experimental Procedures) on several gene sets. Each set contains genes that are "similar" to each other for different reasons, for example genes with similar expression patterns or genes from the same pathway. We tested whether HIGHFLY could identify the correct members of each set by leaving out one gene at a time and calculating the similarity between the left-out gene and the rest of the set. We found that HIGHFLY ranks highly the left-out genes when at least one data source holds the information that this gene is related to the remainder of the gene set (for example, the expression data source is informative for the expression-related gene set) (Figure 2). Importantly, regardless of which particular data sources show the strongest performances, the performance of the combined or fused ranking (last column in Figure 2) is highly robust for all sets, it is not influenced by non-informative data sources, and it is almost always greater than 90% compared to a performance of ~50% for randomly assembled sets of genes (Figure 2). These results validate the technical aspects of the implementation and suggest that HIGHFLY performs robust prioritizations on *Drosophila* data sources.

Next, we investigated whether HIGHFLY would be capable of finding genes that interact *in vivo* with *ato*. A training set, called TRAIN_Ato1 was assembled with the following genes: *ato*, *Brd*, *rho*, *Takr86C*, *pnt*, *dpp*, *Egfr*, *da*, *wg*, *sens*, *chn*, and *sca*. Because different sizes and compositions of training sets are possible, we tested the suitability of this training set for ato-related gene prioritization, by performing two tests. First, we assessed the content of some of the trained submodels. The trained GO submodel for this set contains "peripheral nervous system development", "cell fate specification", "eye morphogenesis", "sensory organ development", etc. as highly over-represented terms (p value$<10^{-09}$). The Text submodel contains stemmed terms like "cell fate", "notch", "egfr", "disc". The InterPro submodel has no highly over-represented domains, but "Basic helix-loop-helix dimerization region bHLH" is marginally over-represented (corrected p-value = 0.07). Secondly, we tested the homogeneity of TRAIN_Ato1, by subjecting it to LOOCV and obtained an AUC performance of 98.5%, suggesting that TRAIN_Ato1 is a coherent and internally consistent training set. To test the possibility of obtaining biologically meaningful prioritizations, we performed a pilot test by prioritizing the right arm of chromosome 3 (chr3R) using TRAIN_Ato1 and then divided all the genes on the list into three groups: the top 1/3, the middle 1/3, and the bottom 1/3. From each group the top 30 genes for which stocks with mutant alleles are available from the public stock centers were examined for their modification of *ato*'s proneural activity, using the same bristle induction assay described above. Four positive genes were found in the top group (*rn*, *Antp*, *gro*, and *pros*), none in the middle group, and none in the bottom group (Table 2 and Table S2). Although the power of this preliminary test is greatly limited due to the relatively small number of genes tested (90) and the variability of available alleles, we found these results sufficiently encouraging to proceed with HIGHFLY prioritizations of all twelve modifier loci found in the deficiency screen. However, to further evaluate HIGHFLY, we intentionally chose a less stringent threshold of further validating the top 30% of ranked genes so as to compare

Table 1. HIGHFLY data sources.

Data type	Data source	Training	Scoring
Functional annotation	Gene Ontology [27]	GO term over-representation	Fisher's omnibus
	PubMed abstract profiles	Text-mining using gene-reference relations from FlyBase; average term weight vector	Cosine similarity
	SwissProt keywords [43]	Term over-representation	Fisher's omnibus
	KEGG [44]	Pathway over-representation	Fisher's omnibus
Gene expression	Life cycle of *Drosophila* microarray data [39]	Collection of all the expression profiles of the training genes	Average of 50% best Pearson correlation
	Tissue-specific gene expression in *Drosophila* larvae [40]	Collection of all the expression profiles of the training genes	Average of 50% best Pearson correlation
	In situ expression [45]	FBbt term over-representation	Fisher's omnibus
Protein sequence	InterPro [43]	Domain over-representation	Fisher's omnibus
	BLAST [46]	Ad hoc BLAST database of training genes prot. seq.	Blast test seq. to ad hoc db; rank by e-value
Allele phenotypes	FlyBase records "phenotype manifest in" [27]	FBbt term over-representation	Fisher's omnibus
Genetic interactions and protein-protein interactions	BioGRID [1]	List of training genes and all their interactors	Overlap between the test gene plus its interactors and the training list
	STRING [29]	Idem BioGRID	Idem BioGRID

HIGHFLY training and scoring strategies for each data source.

doi:10.1371/journal.pgen.1000351.t001

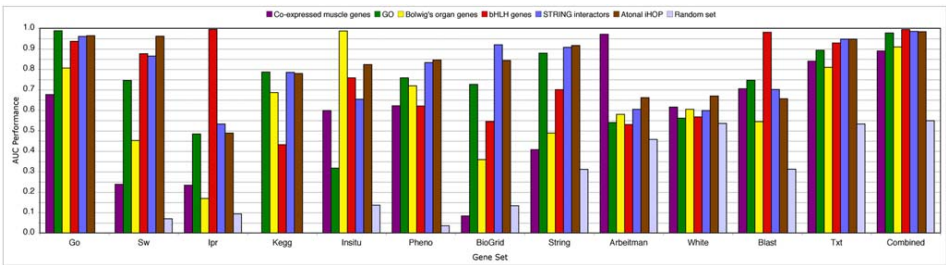

Figure 2. HighFly cross-validation results. The performance values, measured as Area Under the ROC Curve (AUC), obtained for all individual data sources (on the x-axis) are shown for several validation sets (each validation set is plotted in a different color; see legend). The AUC values for the overall prioritization, obtained by integrating all individual rankings, are also shown. Go: Gene Ontology; Sw: SwissProt keywords; Ipr: InterPro protein domains; Kegg: pathway database; Insitu: BDGP *in situ* hybridization data; Pheno: FlyBase mutant phenotypes; BioGrid: genetic and protein-protein interactions; String: protein-protein associations from STRING; Arbeitman: microarray data [39]; White: microarray data [40]; Blast: sequence similarity; Txt: Text-mining PubMed abstracts; Combined: or fused ranking by order statistics. Genes that are functionally related (e.g., same GO annotation or co-occurrence in abstracts) are prioritized well with the GO and text submodels, but also with the STRING and BioGRID submodels. Similarly, prioritization of genetically interacting genes works well with the BioGRID, STRING, GO and Text submodels. Genes that share similar microarray expression profiles or similar *in situ* expression patterns are prioritized well with their respective submodels. Lastly, genes that share similar protein domains are prioritized best by the InterPro and BLAST submodels.
doi:10.1371/journal.pgen.1000351.g002

the rankings of positive and negative genes with a sufficiently large sample size at the end of the screen.

Identification of Novel *ato* Interacting Genes through the Integration of Gene Prioritization and Functional Genetic Modifier Assays

To identify candidate genes within the positive regions, all genes in each of the twelve positive regions were prioritized separately using TRAIN_ATO1 as training set and all 12 HighFly data sources (Table S3). For all genes that were ranked within the top 30%, a mutant stock, when available, was ordered from the public stock centers. Each mutant was then crossed to the sensitized tester fly stock (*uas::ngn*^{bato}*/Cyo;sens,dpp-Gal4/TM6c*) and the bristles at the anterior-posterior margin (where *dpp-Gal4* is expressed) were counted and compared to the number of bristles observed in the control flies as described above. For twelve genes, namely *toc*, *lilli*, *Sbb*, *fj*, *mus209*, *zip*, *shg*, *Egfr*, *dom*, *smg*, *cas* and *ppan*, the number of bristles was significantly lower or higher ($p<0.01$) than in the control flies (Table 2, bottom panel). Each of these mutants were

Table 2. Validation of the HighFly screen results.

Name	Flybase ID	Chromosome	Rank on test region	Rank ratio on test region	Rank on chromosome	Rank ratio on chromosome	Phenotype[a]	P-Value
Antp	FBgn0000095	chr3	44/3341	1.31%	67/6027	1.10%	−31.40%	<0.001
gro	FBgn0001139	chr3	58/3341	1.74%	84/6027	1.40%	−69.90%	<0.001
pros	FBgn0004595	chr3	33/3341	0.99%	48/6027	0.80%	dead	
rn	FBgn0003263	chr3	13/3341	0.39%	21/6027	0.40%	44.20%	<0.001
cas	FBgn0004878	chr3	1/103	0.97%	123/6027	2.00%	−33.60%	<0.001
dom	FBgn0020306	chr2	12/287	4.18%	413/5252	7.90%	−31.40%	<0.002
Egfr	FBgn0003731	chr2	1/287	0.35%	2/5252	0.03%	−51.30%	<0.001
fj	FBgn0000658	chr2	3/118	2.54%	63/5252	1.20%	−50.00%	<0.001
lilli	FBgn0041111	chr2	3/64	4.69%	245/5252	4.70%	−35.40%	<0.002
mus209	FBgn0005655	chr2	2/44	4.55%	484/5252	9.20%	−36.70%	<0.005
ppan	FBgn0010770	chr3	3/58	5.17%	642/6027	10.70%	−33.60%	<0.002
sbb	FBgn0010575	chr2	2/118	1.69%	207/5252	3.90%	−33.60%	<0.001
shg	FBgn0003391	chr2	2/287	0.70%	30/5252	0.60%	−66.40%	<0.001
smg	FBgn0016070	chr3	3/18	16.67%	279/6027	4.60%	−100%	<0.001
toc	FBgn0015600	chr2	1/64	1.56%	834/5252	15.90%	−33.60%	<0.001
zip	FBgn0005634	chr2	1/51	1.96%	123/5252	2.30%	−38.10%	<0.001

Validation of HighFly on prioritized 3R chromosome and on different prioritized deficiency regions.
[a]The average percentage change of the number of ectopic bristles, compared to wild type controls.
doi:10.1371/journal.pgen.1000351.t002

then tested against *uas::sc*; *uas::ngn1*, and *uas::ato* under the control of *dpp-Gal4* to check for the specificity of *ato* interaction. All of the genes modified only the *ato* gain of function phenotype (data not shown). We note that although mutants of genes that ranked in the top 30% of each locus were tested, 11 of the 12 ranked in the top 6% of their locus (Table 2 and Table S3), suggesting that HighFly prioritizations enrich strongly for positive interactions. Similar prioritizations were obtained by using a different high-quality training set (LOOCV AUC = 99.5%), assembled by selecting all 18 known interactors of *ato* from BioGRID (data not shown). In contrast when the same 12 genes were prioritized using 100 randomly assembled training sets, the median rank ratio was 0.247 compared to 0.02 for the *ato* training set (Figure 3).

An alternative analysis, instead of prioritizing each deficiency region separately, is to pool all candidate genes from the positive deficiency regions and prioritize this set in one analysis. We performed such a prioritization as post-analysis and found all 12 positives ranked in the top 10% (Table S1 and Figure S1). An examination of the contribution of individual data sources to the high rankings of the positive genes shows that for all positives, their high ranking is caused by high rankings for several data sources, rather than a single high ranking for one of the data sources (Figure S1), which supports our initial assumption of the added value of data integration for gene prioritization. In a second post-analysis, by comparing HighFly with existing online tools such as FlyBase [27], UCSC Gene Sorter [28], and STRING [29], we found that the use of a training set of genes related to *ato* is more favorable than a single gene query; and also that a gene ranking is more favorable for gene identification than a gene filtering (e.g., using a selection of Gene Ontology terms or a selection of FlyBase expression terms) (Text S1).

Functional inspection of the 16 positive genes (12 from the deficiency screen +4 from the pilot screen of 90 genes on chromosome 3R) by Gene Ontology statistics [30] revealed that this gene set is significantly enriched for developmental processes that require *ato* such as eye development and regulation of transcription (Table 3). Finally, we compared the phenotypic distribution of the effects of the modifier genes identified in our screen with the distribution documented for saturating forward genetic screens and cellular siRNA screens [31]. We find that despite the relatively small number of genes that need to be tested in a HighFly screen, the distribution of phenotypes mirrors that obtained in genome wide forward and reverse genetics screens (Figure 4). These data further support the power and accuracy of the integration of computational biology and genetics.

ato acts as a proneural gene for two different types of founder cells. The first is a subset of sense organ precursor (SOP) of the body wall and appendages and the second is the R8 founder cell of the retina. The major difference between the SOP and the R8 is that the SOP undergoes cell division to generate the sensory organ, whereas the R8 cell terminally differentiates. However, both cells share the property of recruiting neighboring cells into the *ato*-dependent fate; a property unique to *ato*, not shared by other proneural genes. We assessed whether genes identified in one context, also operate in the other. To this end, we tested the relationship between *ato* and its putative interactors in the developing fly retina, where *ato* function is well described [32]. In the retina, *ato* specifies the first photoreceptor, or R cell, the R8 (Figure S2A, B). The R8 then releases an *ato*-dependent EGF signal that organizes the rest of the retinal field and specifies the R1–R7. Loss of *ato* function in the retina results in the complete failure of retinal specification [33]. Expression of an *ato-RNAi* construct (A kind gift of A.P. Jarman) in the eye in *ato* heterozygous flies (*uas::ato-RNAi;h-Gal4*, ato^{1}; see Materials and Methods) reduces R8 specification and consequently the recruitment of other R cells in a dose dependent fashion (Figure S2C,D). One copy of *ato-RNAi* produces a smaller eye with approximately half the normal number of ommatidia (Figure 5A,B). Mutants for the 16 genes identified in the screen were crossed to the *ato-RNAi* flies and scored for their ability to dominantly modify the *ato* RNAi phenotype. Ten of the 16 tested genes, namely *gro*, *rn*, *EGFR*, *cas*, *ppan*, *toc*, *sbb*, *fj*, *shg* and *dom* dominantly enhanced the *ato-RNAi*

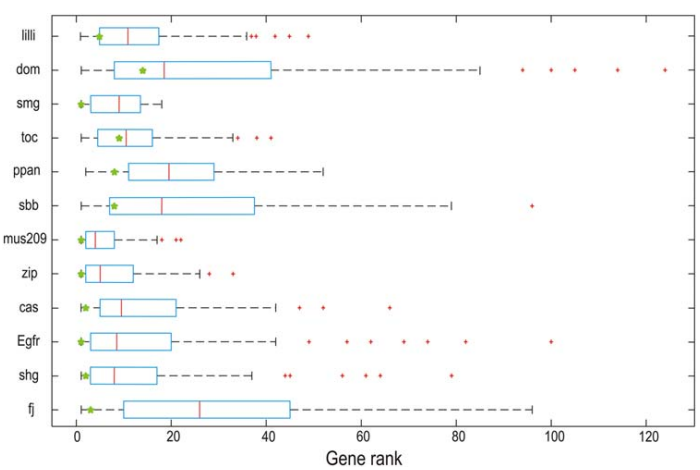

Figure 3. Ranking specificity of the *ato* interacting genes in the bristle assay. The observed rank of a positive gene, using the Atonal specific training set is compared to its rank obtained with a random training set (100 times). Shown is a boxplot of the 100 rankings of each positive gene using random training sets (y axis). The green asterisk represents the rank of the positive for the Atonal training set.
doi:10.1371/journal.pgen.1000351.g003

Table 3. Enrichment of GO-terms among the positive genes of the screen.

GO ID	Genes	Group count	Total count	P value	GO term
GO:0007444	*rn; pros; lilli; fj; ppan; EGFR*	6	259	0.000214	imaginal disc development
GO:0046530	*pros; lilli; sbb; EGFR*	4	72	0.000255	photoreceptor cell differentiation
GO:0016477	*zip; shg; dom; sbb; EGFR*	5	160	0.000276	cell migration
GO:0000904	*pros; lilli; shg; sbb; EGFR*	5	162	0.000284	cell differentiation
GO:0003700	*rn; cas; pros; lilli; Antp; sbb*	6	314	0.000447	transcription factor activity
GO:0007417	*cas; pros; shg; EGFR*	4	95	0.000622	central nervous system development
GO:0007420	*cas; shg; EGFR*	3	34	0.000693	brain development
GO:0007560	*rn; pros; lilli; fj; EGFR*	5	210	0.000764	imaginal disc morphogenesis
GO:0035218	*rn; fj; EGFR*	3	37	0.000811	leg disc development
GO:0001745	*pros; lilli; fj; EGFR*	4	108	0.000811	compound eye morphogenesis
GO:0007164	*fj; zip; EGFR*	3	38	0.000811	establishment of tissue polarity
GO:0000278	*zip; ppan; toc; mus209; EGFR*	5	223	0.000811	mitotic cell cycle

Selection of enriched GO-terms across the 16 positive genes. Full results table is available at http://med.kuleuven.be/cme-mg/lng/HighFly. All genes from chr2 and chr3 are used as background set.

doi:10.1371/journal.pgen.1000351.t003

phenotype, with nine showing further reductions in eye size to approximately 250 ommatidia (Figure 5C, D). An 11th gene, *pros*, was semi lethal. The remaining five genes, namely *Antp*, *smg*, *lilli*, *mus209*, and *zip*, did not appear to alter the *ato-RNAi* induced small eye size.

The data thus far suggest that at least 10 of the 16 genes we identified in the sensory bristle screen also interact with *ato* during retina development. Some of these genes such as *pros* are known for their role in neurogenesis [34], while the EGF receptor is well known for its close interactions with *ato* [35]. However, most of the genes we identified as genetic interactors of *ato* have not, to our knowledge, been previously shown to play a role in *ato*-dependent neurogenesis. Next we asked if these genes might be co-expressed with *ato* in the various PNS anlagen that derive from *ato* expressing precursors. We were able to obtain LacZ enhancer trap lines from stock centers for 10 of the 16 interacting genes (*dom*, *fj*, *lilli*, *mus209*,

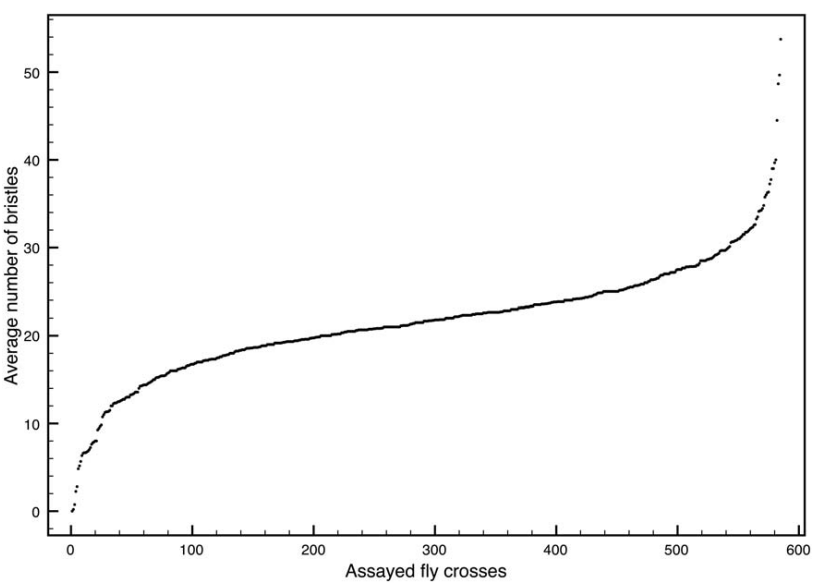

Figure 4. Distribution of the phenotypic range of the average number of bristles per genotype (n>10) plotted for all ~600 assayed genotypes (y-axis). The shape of the curve conforms to the expectations for quantitative screens [31].

doi:10.1371/journal.pgen.1000351.g004

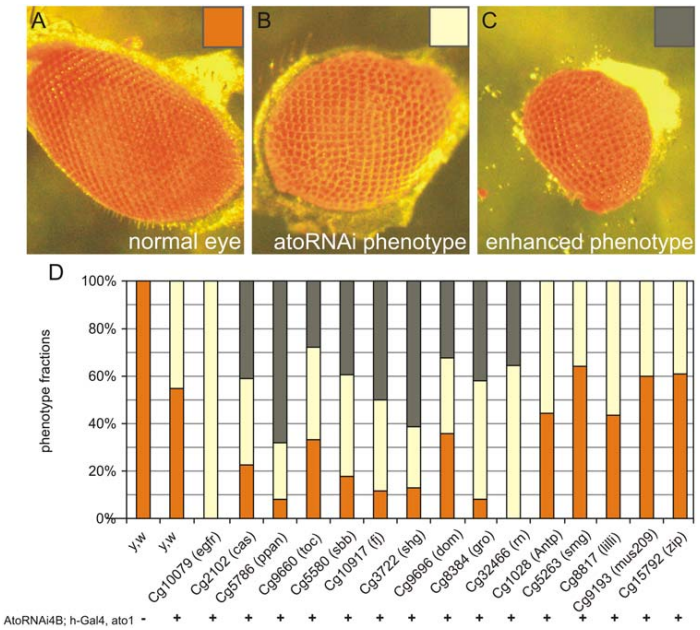

Figure 5. Effect of modifiers on eye size in atonal sensitized eyes. (A) Wild type eye with around 800 ommatidia. (B) effect from *ato-RNAi* on amount of ommatidia resulting in a population of flies with about 400 ommatidia per eye. (C) enhancement of *ato-RNAi* phenotype resulting in a population with smaller eyes with around 250 ommatidia per eye. (D) Overview of phenotypes observed when modifiers were crossed with atoRNAi flies, removal of one copy of the modifier results in a larger population of flies with smaller eyes for 9 of the previously identified modifiers. The other 5 modifiers did not show an alteration in eye phenotype, compared to the *ato-RNAi* phenotype in controls.
doi:10.1371/journal.pgen.1000351.g005

pros, *rn*, *sbb*, *shg*, *toc* and *zip*) to examine their expression patterns in the third instar larval (L3) imaginal discs. In the eye, antennal, leg and wing L3 discs, Ato marks the progenitor pools and the very early precursor cells of specific neuronal lineages. Senseless then marks the precursor cells during and after Ato expression. One enhancer trap, *rn*, did not show any obvious expression relationship to *ato*. Two of the 10 genes, *mus209* (fly PCNA) and *sbb* are generally expressed. An additional two lines, *toc* and *zip* showed expression in the posterior part of the eye disc (Figure S3A), suggesting a later function than that of *ato*. Finally, other five of the 10 tested enhancer traps showed a clear expression relationship with Ato (Figure S3). We observed strong lacZ expression in the L3 discs in Ato-expressing and Ato-dependent cells in the eye disc (*fj*, *lilli*, *shg*, *pros*), in the antennal Johnston organ precursor cells (*dom*, *shg*, *pros*), in the chordotonal organ precursor cells of the wing and leg imaginal disc (*dom*, *shg*, *pros*) (Figure S3 and data not shown). It should be noted that enhancer trap lines might reflect only part of the total expression pattern of the trapped gene.

Identification of Uncharacterised *ato* Interacting Genes

The data above support the feasibility of rapidly and accurately identifying gene function through the fusion of *in silico* gene prioritization and *in vivo* genetic screens. One issue that faces all gene prioritization approaches is an expected bias towards genes with at a large amount of pre-existing information in several databases. Although this is still valuable in assigning novel functions to known genes, we reasoned that it would be interesting to test the performance of HighFly in the prioritization of genes about which there is little explicit information. Genes with limited annotations can potentially be ranked high due to data sources that are independent of existing knowledge, such as sequence similarity, protein domains, gene expression data, or protein-protein interaction data from high-throughput experiments. Indeed, 30 out of 96 genes, known only by their CG numbers, ranked in the top 10% of the *ato*-specific deletion loci identified in the initial bristle screen (Table S4). The recent availability of a genome-wide *in vivo Drosophila* RNAi library [36] allowed us to test these genes for their interaction with *ato*.

When no off-target effects were predicted, available RNAi lines were ordered and crossed to the *ato-RNAi* flies driven by the *h-Gal4* driver in an *ato* heterozygous background (*uas::ato-RNAi;h-Gal4*, ato^1), as well as two different control lines; *h-Gal4*, ato^1 and *h-Gal4* alone. To avoid potential artifacts resulting from the RNAi approach, we set relatively stringent criteria: we searched for genes that show synthetic lethality specifically and only in combination with *ato-RNAi*, but show no phenotype under the two control conditions.

We were able to obtain a total of 36 RNAi lines for 24 uncharacterized genes ranking in the top 10% of positive deficiency regions. Eleven RNAi lines were lethal under all conditions and could not be evaluated further. The 25 remaining

RNAi lines allowed us to perform knockdown of 17 genes. Of these, 2 genes (*CG1024*, *CG1218*,) caused lethality only in combination with *atoRNAi*, but not under control conditions (Table 4). As a further confirmation for the specificity of these interactions, we tested 51 RNAi lines for the bottom 10% ranking genes in each deficiency. None of these lines showed specific synthetic lethality in combination with *atoRNAi* (data not shown). Thus, the combination of HighFly prioritization, the RNAi library and genetic screening allows the rapid functional identification of previously uncharacterized genes.

An Atonal Interaction Network

The combination of forward and reverse genetics tools and computational biology allowed the identification of 18, mostly novel, genetic interactions with the proneural gene *ato*. We sought to determine if the identified positive genes are functionally associated with each other, with *ato*, and with any of the other training genes that were used originally to identify these genes. To this end we used the STRING [29] protein-protein association predictions at 0.8 confidence level and determined the optimally connected sub-network that can be formed among the 18 positive genes, via maximally two other proteins (see Materials and Methods). We find that a network can be constructed that includes 12 of the 16 known genes (data not shown). As expected, the 2 unknown genes play no role in this analysis because of the lack of STRING data at this high confidence level. This analysis discovers Ato itself as member of the best network that connects the positive genes. We found that the maximal confidence level at which Ato is still part of the network is 0.842, and therefore used this stringency for further analyses. The network formed by the 16 known genes at this confidence level (Figure 6A) contains 84 nodes and 250 edges, and now includes 12 of the 16 positive genes and 6 training genes, including *ato*. *Egfr* is directly connected to ato; *fj*, *Antp* and *gro* are connected to *ato* via one other protein; *pros*, *rn*, *shg*, *lilli* are connected to *ato* via two other proteins and *cas*, *smg*, *zip*, *mus209* via three other proteins.

To determine the significance of finding a large interconnected network, which includes *ato*, starting from the 16 positive known genes, we generated 1000 random sets of 16 known genes. Specifically, we used only genes with a name in FlyBase and at least one GO biological process annotation. Only 29 of the 1000 networks contain *ato* and, on average, they contain 0.70 (S.D. = 1.13) training genes, 7.83 nodes (S.D. = 9.09), and 13.07 edges (S.D. = 19.28). An example of such a network is shown in Figure 6B. With a p-value of 0.029 to find Ato in the real network, $p<0.001$ to obtain 84 nodes, $p<0.001$ to obtain 250 edges, and $p = 0.001$ to recover 6 of the 11 training genes, we conclude that the positive genes we identified are strongly associated with each other and with Ato and its known interactors.

A Database of Genome-Wide Gene Prioritizations in *Drosophila* for Ten Canonical Signalling Pathways

A particular feature of the HighFly tool is the speed of prioritization. We wondered whether this computational efficiency makes it possible to prioritize whole chromosomes or even the entire genome. To this end we asked if it is possible to rank the 16 known genes identified in our screen on their respective chromosomes, and if so, whether these rankings would be high. Table 2 shows the chromosomal rankings of these genes. All except one of the known genes rank within the top 10% of their respective chromosome (Table 2).

These data suggest that it is possible to obtain strongly meaningful gene prioritizations across large data sets. We sought to illustrate the general applicability of fly gene prioritization and simultaneously generate a second community-wide resource by prioritizing the entire genome to identify genes that are related to, or potentially involved in, either of ten signaling pathways, namely Transforming Growth Factor beta (TGFβ) receptor signaling pathway (GO:0007179), Epidermal Growth Factor Receptor (EGFR) signaling pathway (GO:0007173), Fibroblast Growth Factor Receptor (FGFR) signaling pathway (GO:0008543), Notch (N) signaling pathway (GO:0007219), Sevenless (Sev) signaling pathway (GO:0045500), Smoothened/Hedgehog (Smo/H) signaling pathway (GO:0007224), Toll signaling pathway (GO:0008063), Extracellular signal-Regulated Kinase (ERK; GO:0007259), JAK-STAT (GO:00016055) and Wnt signaling pathway (GO:0016055). To investigate the rankings in terms of biological processes we calculated GO over-representations for each top 100 ranked genes, excluding the training genes. We also excluded genes that were ranked in the top 100 for more than two pathways and GO-terms that were over-represented in more than four pathways. We find that typical overrepresented functions are cell adhesion and photoreceptor fate commitment for EGFR-related genes; cell migration for FGFR; neuroblast fate determination and equator specification for Notch; defense response for Toll; and ectoderm development for Wnt, suggesting that the prioritizations are biologically meaningful. Finally, we compared prioritizations for 4 of the 10 pathways- namely ERK, Wnt, Hh and JAK-STAT- for overlap with published genome-wide siRNA screens. We find significant overlap between the top 10% of the genome as prioritized by HighFly and the genes scored as positives in these screens for 3 of these pathways (Figure 7). Only the Hh pathway screen shows poor overlap with the prioritizations. Prioritizations and functional analyses, as well as the HighFly software, are available at http://med.kuleuven.be/cme-mg/lng/HighFly.

Discussion

The molecular unraveling of biological processes in the post-genome era is characterized by the use of high-throughput experiments and the integration of prior knowledge (e.g., the use of GO-statistics to select microarray generated gene clusters), and is therefore supported and guided by bioinformatics. Genetic screens in model organisms such as *Drosophila melanogaster* are also high-throughput experiments, but they are yet to be aided by computational techniques, as an integral part of the screen itself. We sought to demonstrate the power of an integrated approach

Table 4. Synthetic lethal modifiers of *ato-RNAi* among the unknown genes.

Clone	CG number	Rank	CG-RNAi+atoRNA ;hGal4, ato[1]	CGRNAi+hGal4, ato[1]	CGRNAi+hGal4
18597	*CG1024*	7.7%	Lethal	No Effect	No Effect
31685	*CG1218*	6.6%	Lethal	No Effect	No Effect

Results of the phenotypes observed from RNAi screen.

doi:10.1371/journal.pgen.1000351.t004

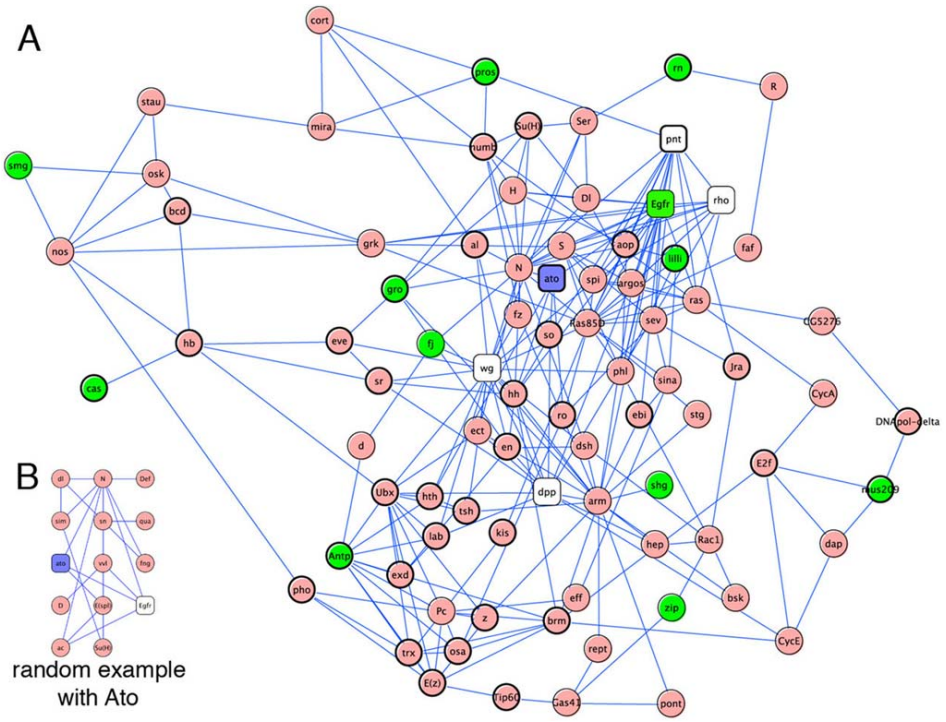

Figure 6. Protein-protein association subnetworks. The subgraphs are extracted from STRING connections with confidence score above 0.842, and aim to connect as many seed genes as possible. Seed genes may be connected through maximally two edges from either side. (A) The 16 positive 'known' genes are used as seed genes to validate their potential relationship. The resulting network is significantly larger and more interconnected than expected by chance and recovers Ato itself as member of the sub-network. (B) Example of a subgraph generated from a random selection of seed genes that does recover Ato. Only 29 out of 1000 random networks recover Ato by chance. These networks are significantly smaller than the network formed by the 16 genes from the screen. Green nodes are positive genes from our screen, while all other nodes are drawn from STRING interaction data (square nodes were part of the training set).
doi:10.1371/journal.pgen.1000351.g006

that combines high-throughput *in silico* and *in vivo* genetic approaches. This integration allowed us to quickly identify novel genetic interactions during neural development in the fly PNS, while significantly reducing the workload of the genetic screen. First, a classical deficiency modifier screen is performed. Then, instead of assaying all the genes located within the positive deficiency regions, the best candidates are selected computationally. This is done by integrating multiple heterogeneous genome-scale data sources, both representing published knowledge (e.g., functional gene annotations or protein-protein interactions), genome sequences, and experimental data (e.g., gene expression data or phenotypes). As such, we were able to assign novel functions for known genes whose involvement in *ato*-dependent neural development was unknown, as well as describe functions for uncharacterized genes.

A major advantage of genetic screens is that they are unbiased: they can reveal a function for a previously unknown gene. Although gene prioritization based on available data would have been expected to affect this property of screens, our data indicate that this is not necessarily the case. Even genes with very little explicit information, and no known function could be identified both as high ranking and as *bone fide* interactors *in vivo* in our HighFly supported screen. In addition, our data suggest that the combination of HighFly prioritizations and transgenic RNAi lines can result in very rapid functional gene discovery.

The use of an integrative screening strategy combining computational biology with medium or high-throughput screening assays is likely to be applicable to a broad range of screening assays (from *in vitro* to *in vivo* assays) beyond *Drosophila* genetics. Essentially any assay designed around evaluating a given gene, and for which whole-genome screening is outside the reach of the typical lab, could benefit from strategies similar to ours. Even with more extensive resources, it may be more productive (at equal time and cost) to evaluate several prioritized screens than a single whole-genome screen. Obviously, the strategy we propose is not applicable in the case where extremely little is known about the molecular basis of a phenotype (because of lack of a training set) while a genetic screen would still be feasible. It is a clear research

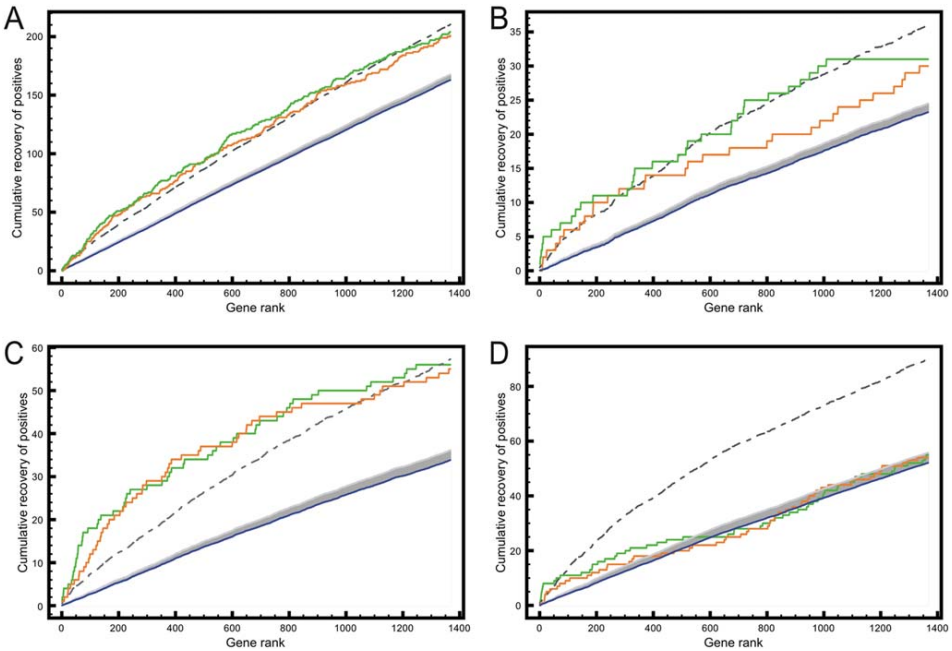

Figure 7. Comparison of whole-genome prioritizations for signaling pathways with results from RNAi screens obtained from http://www.flyrnai.org. For each of the four prioritizations the training set is based on Gene Ontology annotation for the respective pathway, namely GO:0000165 (MAPK) for the ERK pathway (A), GO:0007259 for JAK-STAT (B), GO:00016055 for Wnt (C), and GO:0007224 (smoothened) for hh (D). The green curve represents the cumulative recovery of positive genes when moving down the top 10% of the ranked gene list, using the full training set. The orange curve is similar to the green curve, but now excluding the known GO-annotated positives from the RNAi screen from the training set. The blue control curve is the average recovery curve of the positives, using 100 random training sets of known GO-annotated genes. The grey area represents a 95% confidence interval above the mean and the dotted curve represents two standard deviations above the mean, so that every point above the dotted curve represents a significant ($p<0.05$) enrichment of true positives. For three out of four, namely ERK, Wnt and JAK-STAT, a significant enrichment of positives is found at nearly all thresholds in the top 10% of the genome.
doi:10.1371/journal.pgen.1000351.g007

challenge for computational biology to develop methods applicable to such a situation.

A further advantage of our integrated systems genetics approach is the combination of speed and accuracy of gene function discovery. In this work we tested a total of 180 deletion lines, 220 mutants and 36 RNAi lines to identify 18 *ato* interacting genes, representing a discovery rate of ~5%. It should be noted that the 220 mutants tested include 90 mutants examined only for the purposes of testing the prioritizations as well as 78 mutants ranking between 10% and 30% of their deletion regions. Our data clearly indicate that testing genes ranked in the top 10% only will suffice to discover the vast majority of sought after genes: 17 of the 18 genes identified (~94%) rank in the top 10% of their tested regions. Thus, assuming all genes have available RNAi lines or mutant alleles, testing only 96 genes, after the initial deficiency screen, would have identified at least 17 *ato* interacting genes, a discovery rate of almost 18%. In this regard we note that ENDEAVOUR-based prioritizations appear to outperform existing tools. We believe this to be due to three main properties namely the use of a multi-gene training set, the integration of multiple data sources, and the production of gene rankings.

The genes we find to interact with *ato* reveal an interaction network underlying early neural differentiation. Network analysis reveals two important aspects of the screen. Although neither Ato nor its known interactors were included in the query, the best network found includes Ato and almost all of its known interactors. In addition network analysis yields a number of interesting insights. First, most of the 89 genes in this network are signaling molecules and transcription factors belonging to the Notch, Wnt, EGFR, Dpp and Hh pathways. These pathways are known to interact with *ato* and our data suggest that the newly identified *ato* interacting genes may be members of these pathways or may implement the interactions between *ato* and these pathways. Second, most of the genes tested for both bristle formation and retinal development interact with *ato* in both assays. This suggests that *ato* may work with a core group of genes to implement context-specific neural fate decisions. One exception to this appears to be genes acting in cell division (*mus209*, *lilli*, *zip*) that, not surprisingly, interact in the bristle assay, but not the R8 assay. Third, we note that HIGHFLY was able to predict the interaction of uncharacterized genes with *ato*, which network analysis alone, would have not been able to predict.

In summary, a systems genetics [37] approach not only identifies novel functions for individual genes with great speed and accuracy, but, as would be desirable in a systems biology context, also uncovers the structure and functional attributes of the network formed by these genes. Yet, the main advantage of systems genetics over other systems biology approaches is that the results are physiologically relevant by definition, because they are discovered directly *in vivo*.

The HIGHFLY tool can perform prioritizations on the entire fly genome. We have done this for ten major signaling pathways, but many other prioritizations are possible, depending on the interest of the user. HIGHFLY and its prioritizations are public resources that we hope will contribute to enhancing the speed and accuracy of functional gene discovery *in vivo* and establishing classical genetics as a fundamental tool of systems biology.

Materials and Methods

Fly Strains and Genetics

All crosses were performed at 25°C, except for the *atoRNAi* eye screen crosses which were performed at 28°C, on standard fly food. Deficiency kits, *LacZ* enhancer trap flies and all mutant lines were obtained from the Bloomington and Szeged stock centre. The *atoRNAi* lines were kindly provided by Andrew Jarman, and the RNAi lines for uncharacterized genes were obtained from the Vienna *Drosophila* RNAi Center (VDRC).

Immunohistochemistry

Third instar larval imaginal discs were dissected in 1× PBS. Discs were fixed with 4% formaldehyde in 1× PBT for 15 minutes. Then, washed five times (15 min/T) in 1× PBT. Blocking and antibody incubation were performed as described [38]. The antibodies used were: sheep anti-ATO (1:250), rabbit anti-GFP (1:1000), rat anti-Elav (1:100), guinea pig anti-SENS (1:1000) mouse anti-βgal (1:1000), rabbit anti–βgal (1:1000). Secondary antibodies were always used 1 in 500. Samples were mounted in Vectashield mounting medium and detected using confocal microscopy (BioRad 1024, Hercules, California, United States and Leica DM-RXA, Wetzlar, Germany).

Genetic Screen

The fly strain *w*; *UAS::ngnbato/CyO*; *sens*, *dpp-Gal4/TM6* was used to set up crosses with deficiency lines. The number of the ectopic bristles was used as a parameter to reflect the strength of the proneural function of Ato in this context [24]. When a deficiency region caused a significant change in the number of ectopic bristles, the corresponding deficiency line was further crossed to three fly lines *UAS::ato*; *dpp-Gal4*, *UAS::ngn*; *dpp-Gal4* and *UAS::sc*; *dpp-Gal4* and the number of ectopic bristles was counted. Deficiencies were considered *ato* specific when they altered the amount of bristles generated by *UAS::ato/cyo*; *dpp-Gal4/TM6*, and not by *UAS::ngn*, *dpp-Gal4/TM6* or *UAS::sc*, *dpp-Gal4/TM6*. Within these deficiency regions, high-ranking mutant lines available in the stock centre were ordered and crossed to *w*; *uas::Ngnbato/CyO*; *sens*, *dpp-GAl4/TM6*. If a mutant still caused a significant change in bristle number, the corresponding gene interacts with Ato. The positive genes were tested with flies expressing *UAS::ato*, *UAS::ngn1* and *UAS::sc* respectively under *dpp-Gal4* control to check for specificity. All ectopic bristles were counted under stereomicroscope. For all statistic analysis, the sample number is n = 10, and a significant difference between two average values is defined as $p \leq 0.01$. The eye phenotype screen was performed by crossing *w*; *UAS::atoRNAi/CyO*; *h-Gal4*, *ato^{1}/TM6C*, which reduced the eye size in 50% of the flies, with the mutant strains identified in the bristle screen. Positive genes for retinal modifiers of *ato* were mutants that enhanced or suppressed the atoRNAi phenotype. RNAi strains were crossed to *h-Gal4*; *ato^{1}/TM6* and *h-Gal4* as controls. Only the one showing synthetic lethality specifically with *w*; *UAS::atoRNAi/CyO*; *h-Gal4*, *ato^{1}/TM6C*, but not with two controls was considered as positive.

Gene Prioritization

The gene prioritization method [3,23] works as follows. First, a set of training genes is defined to describe the particular process under study. For each data source, the following data for the training genes are assembled: (1) a gene's function derived from FlyBase GO annotation, textual information extracted from PubMed abstracts, SwissProt keywords and KEGG pathway membership; (2) a gene's expression pattern derived from two general *Drosophila* microarray data sets [39,40] and embryonic *in situ* expression patterns from the Berkely Drosphila Genome Project (BDGP); (3) a gene's protein sequence from Ensembl and its protein domains from InterPro; (4) described mutant phenotypes from FlyBase; and (5) described genetic interactions or predicted protein-protein associations from BioGRID and STRING. The applied training and scoring strategies for each data source are described in Table 1. For each gene in a "test set" the similarity with a submodel is calculated and the ranks according to individual submodel scores are integrated using order statistics, yielding a q-value. The q-value is transformed into a p-value according to fitted distributions, depending on the number missing values. Finally, the test genes are ranked according to this p-value.

Leave-One-Out Cross-Validation (LOOCV)

We assembled sets of genes involved in the same signaling pathway, tested on eight pathways defined by GO; genes with similar expression patterns using an expression cluster from Arbeitman et al. [39] and a second cluster of all genes expressed in Bolwig's organ from FlyBase; genes with the same protein domain, namely the bHLH domain; all genes that interact with the same gene, tested on all interactors with Atonal from BioGRID; and genes that are co-cited with a specific gene in PubMed abstracts, namely genes cited with *ato*, extracted using iHOP [41]. In LOOCV, every gene from every validation set is, in turn left out, and the ranking of the left-out gene within a set of 99 randomly selected genes is recorded. From all these rankings, Receiver Operating Characteristic (ROC) curves are generated and the area under this ROC curve is used as a measure of the performance of each individual data source and of the integrated prioritization.

Network Extraction

The aim of the network extraction is to obtain a subgraph that connects the genes of interest (the seed genes). Network connections were extracted from the STRING protein-protein associations, using a minimum edge confidence (above 0.8). We define the connecting nodes (the non-seed genes) in the subgraph as the nodes that are on the shortest path(s) between two or more seed genes. To identify those connecting nodes, a multiple sources breadth-first search is performed, which is initialized with the seed genes. During the search, the minimum distance to the seed genes is recorded until seed genes are reachable from one another. Upon completion, the final network is obtained by exploring the shortest paths, starting from the seed genes, that have a maximum length of 4 and that connect at least two seed genes. Hence, the extracted network is made of one or more connected components and may not include all the seed genes. The obtained networks were visualized using Cytoscape [42].

Supporting Information

Figure S1 Contributions of the HIGHFLY data sources to the overall ranking. One HIGHFLY prioritization was performed on all 1056 genes that are contained within the 12 positive ato-specific deficiency regions. The first column shows the rank of the positive genes for the overall ranking obtained by the fusion of all the individual sources (columns 2–13). Grey squares represent missing data for that particular gene and data source. The genes with no or limited existing knowledge, such as CG1218 and CG1024 can still be ranked high. CG1218 is ranked high because of similarities with the training set through BioGRID (CG1218 interacts with *sine oculis*), BLAST (CG1218 has sequence similarity with *chn*, E-value 18.2) and Microarray_2 (similarities between CG1218 and the training set according to microarray gene expression data). CG1024 has similarities with the training set through BLAST (CG1024 has sequence similarity with *senseless*, E-value 0.54), InterPro (CG1024 contains a Zinc finger motif, C2H2-type, like *senseless*), and Swissprot (CG1024 contains the keyword "Zinc-finger, DNA binding").
Found at: doi:10.1371/journal.pgen.1000351.s001 (0.51 MB TIF)

Figure S2 Expression of *ato-RNAi* inhibits retinal differentiation. Eye discs are oriented with posterior located to the left. A) Scheme of *ato* dependant retinal induction using the formation of one photoreceptor cluster as example, first Ato is expressed in a stripe of cells, than, due to lateral inhibition *ato* expressing cells are restricted to three cells and then a single cell, the R8, which begin to express Sens. The other 7 photoreceptors are recruited in a reiterative way. When these neurons mature they express Elav. B) wild type control eye disc stained for Elav (blue), Sens (green) and Ato (red). C, D) Expression of *ato-RNAi* causes dose-dependent loss of retinal differentiation with one copy (C) leading to the appearance of gaps in the Elav pattern, and two copies (D) leading to a major failure of photoreceptor differentiation.
Found at: doi:10.1371/journal.pgen.1000351.s002 (0.93 MB PDF)

Figure S3 Overview of the expression pattern of *ato* interacting genes detected using *LacZ* enhancer trap lines or antibodies. (A) Overview of eye-antenna imaginal disc, with posterior to the right, ed: eye disc, ad: antenna disc. (A′) β-gal staining of *LacZ* enhancer trap flies mimicking the expression pattern of the genes nearby. Enhancer traps of *zip*, *ff*, *sbb*, *shg*, *toc*, and *lilli* show expression patterns in the eye disc. (B) eye disc of *lilli* enhancer trap flies, showing co-localization between Ato (B″) Sens (B′) and β-gal (B′). (C) Eye disc of *ff* enhancer trap flies, showing co-localization between Ato (C″) Sens (C′) and β-gal (C′). (D) Leg disc of *shg* enhancer trap flies, showing co-localization between Ato (D″) Sens (D′) and β-gal (D′) in the leg chordotonal organ precursor. (E) Wing disc of *dom* enhancer trap flies, showing co-localization between Ato (E″) Sens (E′) and β-gal (E′) in the wing chordotonal organ precursor. (F) Antennal disc of *dom* enhancer trap flies, showing co-localization between Ato (F″) Sens (F′) and β-gal (F′) in the Johnston organ precursor.
Found at: doi:10.1371/journal.pgen.1000351.s003 (3.02 MB PDF)

Table S1 Results of the prioritization of all 1056 genes from the 12 positive deficiency regions together, indicating the result of the bristle assay. Negatives and positives are indicated with orange and green shading respectively.
Found at: doi:10.1371/journal.pgen.1000351.s004 (0.38 MB XLS)

Table S2 Results of the prioritization of chromosome 3R. The first 30 available mutant stocks are shown for the top 1/3, middle 1/3, and bottom 1/3 of chromosome 3R after prioritization. Negatives and positives in the bristles assay are indicated with orange and green shading respectively.
Found at: doi:10.1371/journal.pgen.1000351.s005 (0.04 MB XLS)

Table S3 Results of the prioritization (Dec 2005) of the 12 positive deficiency regions, in 12 sheets, indicating the mutant alleles tested in each region and the result of the bristle assay. Negatives and positives are indicated with orange and green shading respectively.
Found at: doi:10.1371/journal.pgen.1000351.s006 (0.16 MB XLS)

Table S4 Results of the prioritization (June 2007) of the 12 positive deficiency regions, in 12 sheets, indicating the RNAi lines tested in each region and the result of the bristle assay. Negatives and positives are indicated with orange and green shading respectively.
Found at: doi:10.1371/journal.pgen.1000351.s007 (0.58 MB XLS)

Text S1 Supplementary Analysis: comparison of HIGHFLY with existing tools through post-analysis.
Found at: doi:10.1371/journal.pgen.1000351.s008 (0.12 MB PDF)

Acknowledgments

Special thanks to Robin Hiesinger for stimulating discussions. We thank H. J. Bellen, the Developmental studies Hybridoma Bank and the VDRC for flies and antibodies. We especially thank A. P. Jarman for sharing the unpublished *ato* RNAi lines and Juan Modolell for the *UAS::sc* flies.

Author Contributions

Conceived and designed the experiments: SA LCT RB YM BAH XJQ. Performed the experiments: SA SV SH LCT RB JY XJQ. Analyzed the data: SA SV SH LCT RB YM BAH XJQ. Wrote the paper: SA SV YM BAH XJQ.

References

1. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: D535–539.
2. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics 22: 773–774.
3. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24: 537–544.
4. Freudenberg J, Propping P (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18 Suppl 2: S110–115.
5. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, et al. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res 34: e130.
6. Lopez-Bigas N, Ouzounis CA (2004) Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32: 3108–3114.
7. Ma X, Lee H, Wang L, Sun F (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. Bioinformatics 23: 215–221.
8. Oti M, Snel B, Huynen MA, Brunner HG (2006) Predicting disease genes using protein-protein interactions. J Med Genet 43: 691–698.
9. Perez-Iratxeta C, Bork P, Andrade-Navarro MA (2007) Update of the G2D tool for prioritization of gene candidates to inherited diseases. Nucleic Acids Res.
10. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, et al. (2006) TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res 34: W285–292.
11. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, et al. (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res 34: 3067–3081.

12. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, et al. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 33: 1544–1552.
13. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, et al. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res 33: W758–761.
14. Guillemot F (1999) Vertebrate bHLH genes and the determination of neuronal fates. Exp Cell Res 253: 357–364.
15. Campuzano S, Modolell J (1992) Patterning of the Drosophila nervous system: the achaete-scute gene complex. Trends Genet 8: 202–208.
16. Anderson DJ (1999) Lineages and transcription factors in the specification of vertebrate primary sensory neurons. Curr Opin Neurobiol 9: 517–524.
17. Brunet JF, Ghysen A (1999) Deconstructing cell determination: proneural genes and neuronal identity. Bioessays 21: 313–318.
18. Jan YN, Jan LY (1994) Neuronal cell fate specification in Drosophila. Curr Opin Neurobiol 4: 8–13.
19. Cabrera CV, Alonso MC (1991) Transcriptional activation by heterodimers of the achaete-scute and daughterless gene products of Drosophila. Embo J 10: 2965–2973.
20. Hassan BA, Bellen HJ (2000) Doing the MATH: is the mouse a good model for fly development? Genes Dev 14: 1852–1865.
21. Vervoort M, Ledent V (2001) The evolution of the neural basic Helix-Loop-Helix proteins. ScientificWorldJournal 1: 396–426.
22. Quan XJ, Hassan BA (2005) From skin to nerve: flies, vertebrates and the first helix. Cell Mol Life Sci 62: 2036–2049.
23. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, et al. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 36: W377–384.
24. Quan XJ, Denayer T, Yan J, Jafar-Nejad H, Philippi A, et al. (2004) Evolution of neural precursor selection: functional divergence of proneural proteins. Development 131: 1679–1689.
25. Skowronska-Krawczyk D, Matter-Sadzinski L, Ballivet M, Matter JM (2005) The basic domain of ATH5 mediates neuron-specific promoter activity during retina development. Mol Cell Biol 25: 10029–10039.
26. St Johnston D (2002) The art and design of genetic screens: Drosophila melanogaster. Nat Rev Genet 3: 176–188.
27. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, et al. (2008) FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res.
28. Kent WJ, Hsu F, Karolchik D, Kuhn RM, Clawson H, et al. (2005) Exploring relationships and mining data with the UCSC Gene Sorter. Genome Res 15: 737–741.
29. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358–362.
30. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710–3715.
31. Friedman A, Perrimon N (2007) Genetic screening for signal transduction in the era of network biology. Cell 128: 225–231.
32. Jarman AP, Sun Y, Jan LY, Jan YN (1995) Role of the proneural gene, atonal, in formation of Drosophila chordotonal organs and photoreceptors. Development 121: 2019–2030.
33. Jarman AP, Grell EH, Ackerman L, Jan LY, Jan YN (1994) Atonal is the proneural gene for Drosophila photoreceptors. Nature 369: 398–400.
34. Vaessin H, Grell E, Wolff E, Bier E, Jan LY, et al. (1991) prospero is expressed in neuronal precursors and encodes a nuclear protein that is involved in the control of axonal outgrowth in Drosophila. Cell 67: 941–953.
35. Lage P, Jan YN, Jarman AP (1997) Requirement for EGF receptor signalling in neural recruitment during formation of Drosophila chordotonal sense organ clusters. Curr Biol 7: 166–175.
36. Dietzl G, Chen D, Schnorrer F, Su KC, Barinova Y, et al. (2007) A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature 448: 151–156.
37. Hiesinger PR, Hassan BA (2005) Genetics in the age of systems biology. Cell 123: 1173–1174.
38. Mardon G, Solomon NM, Rubin GM (1994) dachshund encodes a nuclear protein required for normal eye and leg development in Drosophila. Development 120: 3473–3486.
39. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, et al. (2002) Gene expression during the life cycle of Drosophila melanogaster. Science 297: 2270–2275.
40. Li TR, White KP (2003) Tissue-specific gene expression and ecdysone-regulated genomic networks in Drosophila. Dev Cell 5: 59–72.
41. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36: 664.
42. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.
43. (2008) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res.
44. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277–280.
45. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, et al. (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3: RESEARCH0088.
46. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

8.2 Contribution of the PhD candidate

The PhD candidate has collected the fruit fly specific genomic data, and has developed the 'Endeavour-HighFly' software. He has also contributed to the computational section of the genetic screen (prioritization of the 12 regions using the atonal mediated gene sets as well as control gene sets). Last point, he has also contributed to the full genome pathway prioritizations whose results have been compared to siRNA screens.

8.3 Discussion

Beside the application described in that paper, we have also applied our gene prioritization approach to several other genetic disorders. The following section describes three applications respectively to Congenital Heart Defects (CHDs) [28, 232] and eye related disorders.

8.3.1 Congenital Heart Defects

The results presented here have been obtained through a close collaboration with the laboratory for the Genetics of Human Development of the Department of Human Genetics, headed by Prof. Koen Devriendt. More precisely, the adaptation of our prioritization strategy for CHD was realized in close collaboration with Bernard Thienpont. This work has been published in the American Journal of Human Genetics [232].

Congenital heart defects are the single most important congenital cause for perinatal mortality and morbidity [100, 233] but despite this manifest importance, their etiology remains largely obscure. Although epidemiological studies demonstrated that certain environmental factors are contributory [114], family and twin studies suggest a major genetic component [45, 151]. Indeed, mutations in several genes were associated with monogenic CHDs, mainly through linkage analysis in large families in which a CHD segregates as an autosomal-dominant trait [190]. Given the mortality associated with CHDs, such large families are rare. Although family studies suggest that partially penetrant causes of CHDs are much more common than purely monogenic causes [42, 257] such loci have remained unidentified in linkage studies. As a result, only few causative genes have been identified, and mutation analyses have shown that they account for only a very small fraction ($< 1\%$) of CHD cases [192], representing a serious limitation in the genetic counseling of CHD patients and their families and in the elucidation of the pathogenesis of CHD. To accommodate these limitations, identification of loci associated with CHDs through chromosomal rearrangements was designed as an alternative strategy.

Such a strategy enables the identification of regions harboring genes involved in heart development in a dosage-sensitive manner and the construction of a human morbidity map for CHDs. Already several candidate loci for CHD that were identified through the screening of patients with a CHD by means of array CGH have been reported [231, 75].

One locus, located on chromosome 6q24-q25, was further delineated and characterized through immunohistochemistry analysis, array comparative genome hybridization, candidate gene prioritization, zebrafish assays, and mutation analysis. The software Endeavour was used to perform the candidate gene prioritization. First, a fine mapping of 11 chromosomal aberrations in 6q24-q25 was realized (six patients with a CHD and five patients without a CHD). This led to the definition of a commonly deleted region (also termed critical region), shared by the 6 CHD patients and none of the control patients. The region is rather small (0,85 Mb) and contains only five genes. The clustering of CHD-associated deletions on 6q24-q25 suggested that haploinsufficiency of one or more genes in this locus causes CHDs. Although this putative CHD gene most likely resides in the commonly deleted region, the alternative hypothesis that two or more genes located elsewhere on 6q24-q25 cause CHDs with incomplete penetrance could not be excluded. The candidate gene prioritization was therefore extended to all genes on 6q24-q25 (105 genes in total).

Endeavour was first adapted to the current problem by adding several novel data sources and by accounting for correlation between the different data sources. The novel data sources are an expression microarray data set of murine heart development (from GEO) and three datasets representing homology data, extracted from HomoloGene [255, 207], BioMart [91, 216], and Inparanoid [170, 174]. This data was summarized by vector representations, and scoring was achieved with the use of Pearson correlation (similar to what is done for expression data). Using multiple data sources in conjunction is expected to reduce the noise and to yield better results. However, there is often redundancy between the data sources, and not accounting for it might biased the analysis towards the knowledge that is redundant. The spearman rank correlation was calculated between any pair of data sources over several heart related prioritizations. Data sources that displayed a moderate to strong correlation (> 0.3) were fused using the order statistics prior to the global fusion, leading to a tree-based prioritization scheme (see figure 8.1.). Additionally, candidate genes were prioritized on the basis of seven distinct training sets (see table 8.1), representing discrete aspects of cardiac development and genetics. Results from the seven sets were fused with the order statistics. Leave-one-out cross-validation demonstrated that this adapted algorithm readily ranks genes with an established involvement in heart development, on average in the top 5%. In addition, and based on this cross-validation, data sources with an AUC below 0.6 were omitted. Gene prioritization yielded a ranked list of candidate genes for CHDs, with MAP3K7IP2 (also known as TAB2) ranking first of all 105

genes from 6q24-q25. Interestingly, it is located in the commonly deleted region. Upon genome-wide prioritization, TAB2 moreover ranks 44th among all human protein coding genes (22742). This tree based structure is also though to reflect more accurately the biology that underlies complex disorders. Indeed these complex traits can result from the perturbation of distinct pathways, each of which can be modeled separately by one branch of the tree.

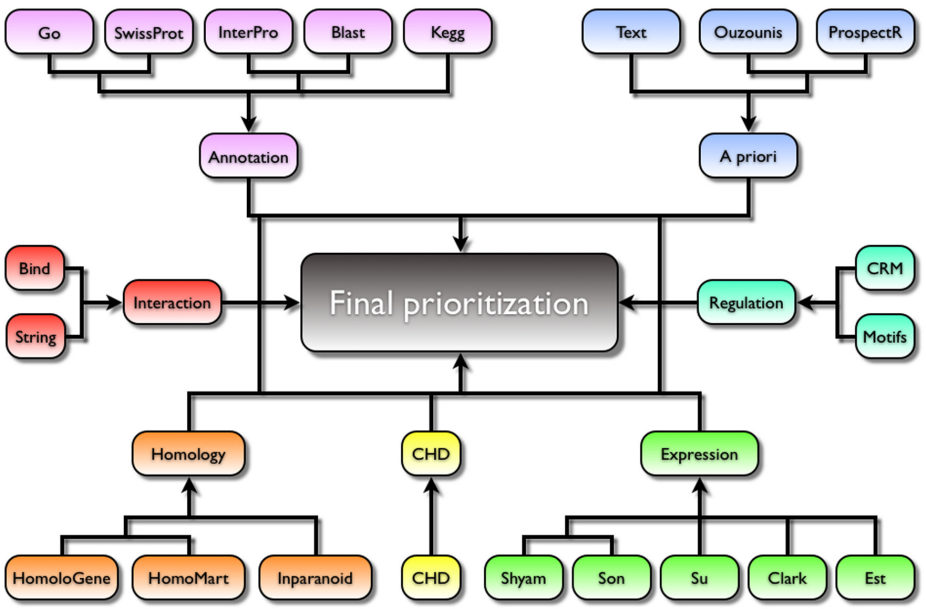

Figure 8.1: Tree based prioritization. The 21 data sources used for the CHD application are depicted by colored boxes that are grouped into seven categories according (i) to the *a priori* knowledge about their content and (ii) to the correlation they exhibit pairwise (spearman rank correlation). These seven categories are then used together to perform the last prioritization. A total of 13 prioritizations is thus needed to obtain the final results. Note: CHD refers to the CHD specific microarray expression dataset.

Immunohistochemistry analysis showed cytoplasmic expression of TAB2 in various part of the developing human heart (ventricular trabeculae, endothelial cells of the conotruncal cushions of the outflow tract, and endothelial cells lining the developing aortic valves). In addition, TAB2 exhibits a compatible expression pattern in zebrafish embryos (cardiac outflow tract, dorsal aorta, posterior cardinal vein). Furthermore, knocked down zebrafish embryos displayed severe heart failure (the heart tube appeared thin and elongated, with blood pooling in the common cardinal vein before entry into the heart) in late development. These results indicate

Biological process	Number of genes	AUC	rank of TAB2
Vascularisation	20	97.98%	5th
Left-right asymmetry	11	92.56%	>5th
Neural crest	11	93.94%	2nd
First heart field	12	93.94%	>5th
Second heart field	12	96.13%	>5th
Valve development	19	93.83%	1st
Known CHD genes	21	93.60%	1st

Table 8.1: The seven CHD specific gene sets used in our CHD application. The sets describes various biological processes that are related to heart development. They have been manually built and are manually maintained by Bernard Thienpont. Their sizes range from 11 to 21, and the estimate of the performance (AUC) of the leave-one-out cross-validation (LOOCV) is always higher than 90%. The rank of TAB2 for which experimental validation has been conducted are displayed (the overall ranking is 1st).

a function of TAB2 in the developing human heart. The role of TAB2 in CHDs was further confirmed by analyzing the DNA sequence of TAB2 in 402 patients with outflow tract defects leading to the discovery of two novel heterozygous missense mutations and by showing that it is disrupted by a balanced translocation in three family members with a CHD. Altogether, these results provide strong evidence that TAB2 has a major role in cardiac development and thus that our candidate gene prioritization method can be tuned and integrated into wet lab workflows to efficiently discover novel disease genes.

8.3.2 Eye disorders

The results presented here have been obtained through a close collaboration with the laboratory for Cytogenetics and Genome Research of the Department of Human Genetics, headed by Prof. Joris Vermeesch. More precisely, the adaptation of our prioritization strategy for eye related disorders was realized with Irina Balikova.

The aim of this research is to identify novel genes that contribute to different eye related disorders such as Usher syndromes or cataracts. The poor number of available patients does not permit the definition of one or several candidate regions such as described previously for CHD. The strategy is therefore to prioritize the human genome and to retain the top several hundreds candidate genes that will be further tested by sequencing on distinct patient cohorts. For a prioritization adapted to the current problem, the integration of novel data sources (SAGE/EST eye data) as well as the development of alternative strategies (filtering step and/or tree-based prioritization) have been investigated and the main results are reported

Gene set	Number of genes	AUC
Cilia	19	97.8%
Usher	8	100%
Cilia and usher	27	98.1%
Mac	27	97.7%
Rho cycle	12	90.2%
Connexins	3	100%
Crystallins	7	100%

Table 8.2: The seven eye disorders gene sets used in our eye disorder application. The sets describes various biological processes that are related to eye development as well as known eye diseases. They have been manually built by an expert, Irina Balikova. Their sizes range from three to 27, and the estimate of the performance (AUC) of the leave-one-out cross-validation (LOOCV) is always higher than 90%. The smallest gene set was excluded from further prioritization to avoid spurious results based on a too small training set. An AUC of 100% (perfect validation) is observed for three of the sets that are very homogeneous and that indeed represent one very specific process or even one single protein complex.

below.

One important feature in this application is that the candidate selection has to be optimal so that a minimum of genes selected via prioritization are false positive genes, that is genes that are not interesting at all regarding eye disorders. One way to control this is to add expression data to make sure that the selected genes are expressed in the eye or in some subpart of teh eye (*e.g.*, lens, retina, cornea, iris, optic nerve). One possibility is the large EyeSage library, from NEIBank and built by Rickman *et al.* [202, 259], that contains tag counts for 51476 distinct tags and for 47 tissues (8 eye related tissues and 39 other tissues). These tags correspond to 12097 genes covering 53% of the protein coding genome. In addition, NEIBank also proposes tissue specific SAGE/EST libraries among which three measure the expression on eye tissue for more than 10000 genes.

This expression data can be used in two different ways. First option, a filter to pre- or post-process the data can be built in order to strictly select the genes expressed in the human eye, ruling out the false positive genes selected by Endeavour but that are apparently not active in the eye. This approach is similar to the *ab initio* methods described in chapter 1. The main advantage of this method is the reduction of the number of candidate from over 22000 to around 12000. However, this is also the main disadvantage as the SAGE libraries are incomplete and a strict filtering approach might also exclude true positive genes. Second option, the data can be integrated within Endeavour as an extra data source similarly to what has been done for CHD. This method has the advantage of being more flexible since the SAGE data influences the results as the other data sources do but is unable

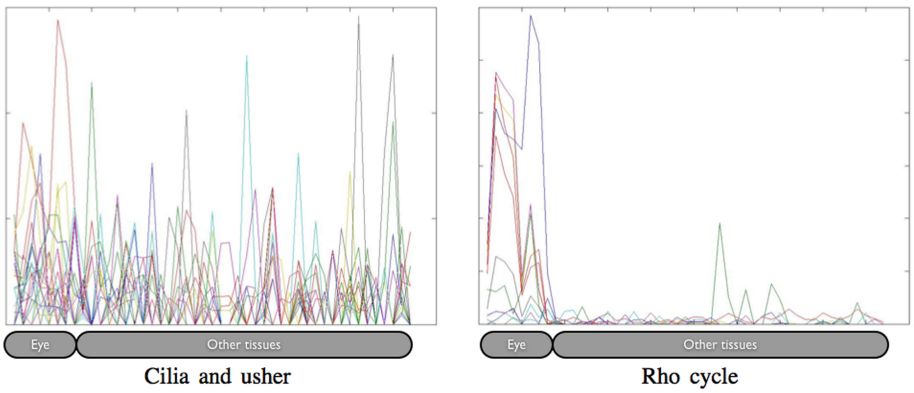

Figure 8.2: Expression profiles of eye disease genes.The expression profile are derived from the SAGE data of NEIBank. The two gene sets related to eye disorders are 'Cilia and usher' together (left) and 'Rho cycle' (right). On the x-axis, the 8 first tissues are eye tissues and are followed by 29 general tissues (from the whole body). It can be observed that for the 'Cilia and usher' set, no pattern can easily be distinguished meaning that these genes are not exclusively expressed in the eye. At contrary, a clear pattern can be distinguished for the 'Rho cycle' for which most of the genes are only poorly expressed in the non eye tissues. Of interest, in both cases, it can be observed that some known genes are shown not to be expressed in the SAGE data illustrating its incompleteness.

to make a decision by itself. It is less conservative, one gene without known eye expression can still be considered as a good candidate with strong evidence from the other data sources. For this second option, two data representation schemes are possible depending on what are exactly the target genes. If the target genes are the genes that are specifically expressed in the eye (and maybe marginally expressed elsewhere), then it is better to use the ratio of the expression in the eye over the expression in the rest of the body. This ratio is very high for tissue specific genes and can be used to rule out the genes that are ubiquitously expressed and that are therefore unlikely to be the cause of non syndromic eye related disorders. Second possibility, the target genes are expressed in the eyes but are not necessarily eye specific, in which case the calculation of the ratio does not make sense. It is then preferable to compute profile similarity to known disease genes on the whole expression profiles that include both eye tissues and other tissues.

In addition, and similarly to the CHD applications, specific training sets have been built and benchmarked. For eye disorders, the seven gene sets cover various eye diseases (usher syndromes, eye cataracts) and key developmental pathways (ciliary muscle development, rho cycle) and are summarized in table 8.2. The benchmark

shows that they are all extremely homogeneous (AUC $> 90\%$) meaning that reliable predictions can be derived from their use in prioritization.

The SAGE data sources and the representation schemes are then examined. The implementation of a pre- or post-processing filter does not represent the best solution since it is excluding more than 20% of the know disease genes (see example in figure 8.2). This is likely due to the incompleteness of the SAGE data and confirms that a filter is too conservative. A benchmark analysis reveals that except for one training set (rho cycle), the SAGE data can not be used to discriminate the known eye disease genes from the genome, an example is given in figure 8.2. Still, the estimated performance indicates that the SAGE data can be used in conjunction with the existing data sources to strengthen the results. The results tend to indicate that none of the two representations performs better than the other, reflecting the fact that tissue specificity can not always be assumed: *e.g.*, usher syndromes main phenotypes are blindness and deafness, which means that expression in the ear system is also expected.

Based on these results, a full genome prioritization was realized for three of the training sets ('cilia and usher', 'mac', and 'rho cycle') using the best performing data sources (including the SAGE data source). These three prioritizations will then be used to determine the candidate genes to be spotted on the sequencing array. Eventually, this array will then be used to screen patients with eye related disorders in order to identify novel disease genes.

8.3.3 CHD wiki

These applications have shown that the scientific community could benefit from prioritization tools. However, and as defined in the introductory chapter, bringing these tools to the researchers themselves is still a challenging task. Chapter 4 of this thesis describes the development of a web based client as an alternative to the Java client. This intuitive and user-friendly web based interface represents a first possible answer to that challenge. The present section describes an alternative strategy in which a gene prioritization module is integrated into a wiki based knowledge database to ease the prediction of novel disease gene based on the existing knowledge. This work was achieved in close collaboration with Roland Barriot, Jeroen Breckpot, Bernard Thienpont, and Sylvain Brohée, and was published in Genome Medicine [28].

In the recent years, many computational strategies have been developed, benchmarked and experimentally validated. However, only few convincing applications with a clear impact on biology have been demonstrated [12], and how to efficiently integrate them into the daily practice of molecular biologists, geneticists, and clinicians remains an open question. There is clearly a gap between such advanced (and somewhat complex) analysis strategies and actual wet lab

work. A similar gap can be observed between those strategies and clinical genetics where increasingly complex molecular data needs to be interpreted towards the diagnosis of constitutional disorders. Beside the complexity, a lack of fine-tuning to the needs of a specific biological research area can handicap a strategy. To bridge this gap and bring integrative analysis strategies into practice, several avenues can be investigated, including workflow managements systems such as Taverna [107], or web based toolboxes such as Galaxy [35, 228]. We introduce another possibility by building on the recent advances in Wiki-based technologies to develop a collaborative knowledge base that integrates a gene prioritization and network browsing portal. We apply this strategy to congenital heart defects and build CHDWiki that aims at mapping genes and genomic regions, and at untangling their relations with corresponding human phenotypes.

The tight integration of our gene prioritization strategy with a CHD knowledge base and its user friendly interface alleviate the problems mentioned above. Practically, the CHDWiki interface is again a simplification of the web client interface that consists of 4 steps: species selection, training gene definition, data source selection, and finally candidate gene definition. CHDWiki is a knowledge base for human congenital heart defects and therefore the species selection step is not necessary. Furthermore, seven CHD related gene sets that represent different aspects of the heart development were built by an expert in the field (Bernard Thienpont, see Table 8.1). This means that, for training, the user only has to select which of these gene sets is suitable, instead of manually inputing the individual genes. These seven gene sets have furthermore been cross-validated in order to assess their homogeneity and to estimate which data sources should be used to perform meaningful prioritizations. This means that, for CHDWiki, the optimal data sources are selected and the user does not need to focus on that. Last step, the user still needs to define the candidate genes to prioritize. Altogether this means that the CHD wiki interface is reduced to two steps: training set selection and candidate gene definition, with the training set selection being much easier than the training gene selection. Last detail, the results of the full genome prioritizations for these seven training sets is calculated beforehand and only the scores for the candidate genes are retrieved on the fly, making the whole prioritization process a lot faster.

Recently, Endeavour was also included in CNV-WebStore, an online platform to streamline the processing and downstream interpretation of microarray data in a clinical context [241]. This platform includes computational tools to preprocess, analyze, visualize and interpret the data. For gene prioritization purposes, Endeavour can be used to prioritize the genes that are found in a CNV associated to a phenotype for instance. This shows that Endeavour, and in the future MerKator, can be integrated efficiently into more complex workflows that integrate different computational tools, and that aim at solving genetics problems.

Data set	Sensitivity		
	at 1%	at 10%	at 30%
Control	$\sim$ 1%	10%	30%
OMIM	46%	83%	95%
MetaCore diseases	33%	75%	92%
GAD	33%	77%	93%
MetaCore pathways	64%	94%	98%
Ingenuity	69%	94%	98%
Gene Ontology	56%	89%	97%
Literature	9%	53%	91%
Literature[a]	29%	71%	100%
Drosophila validation	17%	83%	100%

Table 8.3: The sensitivity (TPR) is indicated for different benchmark datasets and for different threshold values (FPR). The results are presented for three disease benchmark datasets (OMIM, MetaCore, GAD - chapter 7), for three pathway benchmark datasets (MetaCore, Ingenuity, Gene Ontology - chapter 7), for one predictive analysis (Literature - chapter 4), and for the drosophila validation described in the present chapter. [a] Literature analysis with the GWAS study excluded. Results indicate that a threshold of 30% is in most cases enough to retain more than 90% of the positive genes.

8.3.4 Optimal threshold

When a prioritization approach is applied to a real biological question, it is necessary to identify the optimal threshold that will be used to determine the genes to be assayed from the complete ranking of the candidate set. The optimal threshold depends on the cost associated to a false positive and to a false negative. In our setup, the number of positive genes is usually very small (one or a few), and therefore the cost of a false negative is usually very high (*e.g.*, not finding the unique positive gene is costless). At the contrary, the cost of a false positive is usually reasonable (cost of the associated assay) and is furthermore inherent to the unbalance between the positive genes and the negative genes.

The pilot study from the paper described in that chapter indicate that a hard threshold of 30% can be used as a conservative estimate to retain 100% of the positive genes, while still reducing the cost by a factor 3.3. The application of that threshold to the entire chromosomes 2 and 3 revealed that 30% is a conservative estimate since all 12 genes were ranked in the top 15% (and 10 of the 12 were ranked in the top 10%). These results are in agreement with the results of the cross-validation and the predictive studies as shown in table 8.3. Altogether, these results indicate that, although it is difficult to determine a universal threshold, there is evidence that 30% can reasonably be used in practice.

Chapter 9

Conclusion

This thesis focuses on the gene prioritization problem, that can be defined as the identification of the most promising candidate genes, among a potentially large list of genes, with respect to a biological process of interest (*e.g.*, a genetic disease). This thesis introduces two gene prioritization methods, and presents the associated benchmark and validation studies.

We have first developed and implemented a gene prioritization method based on the Order Statistics (OS). It relies on the 'guilt-by-association' concept that is the more promising candidate genes are in fact the ones that are similar to the already identified disease genes. The gene similarity is estimated using multiple genomic data sources, and a consensus is found by adopting a data fusion strategy. The underlying algorithm consists of three steps. In the first step, the genes already known to be involved in the process of interest are used to create a set of models of that process. One model is built per genomic data source used. In the second step, the candidate genes are scored and ranked accordingly per model. The best scores are assigned to the genes that are similar to the model (and therefore indirectly similar to the known genes). In the third step, the rankings from multiple models are fused using the Order Statistics (OS) to obtain a single final ranking. The Order Statistics allow for an efficient handling of the missing data point, which is crucial in our setup. For a single gene, only the positions obtained for data sources for which the gene has data are fused, therefore avoiding a bias towards the well studied genes (that don't have missing data as compared to poorly characterized genes that have missing data). The algorithm is ranking higher, in the final ranking, the candidate genes that are likely to also be involved in the biological process under study.

Our second approach relies on kernel based methods. It uses the same inputs as the OS based method (known disease genes and candidate genes) and also produces

a final ranking of the candidate genes. The core is however different since it is a 1-SVM algorithm that uses kernel matrices that are built from the original data matrices. The use of a kernel based method makes the whole framework more elegant (*i.e.*, easier to extend and to update) and makes the implementation of other kernel based methods straightforward. In addition, it is possible to assign different weights to each data source to reflect its ability to accurately classify the genes. These two prioritization methods were then extended first to support multiple species (*e.g.*, rat, mouse, fly, worm, zebrafish), and second to support cross-species prioritization. In that setup, the methods can prioritize candidate genes from one species using data from many species. This allows the use of model organism specific data to enhance human disease gene discovery.

These methods have been benchmarked through the use of a leave-one-out cross-validation procedure. The leave-one-out procedure mimics the discovery of a single novel disease gene at the time and is therefore conceptually close to the real application setup. The results show that the two methods are effective since we obtain high AUC values (usually greater than 90%). These results stand for different variations of the benchmark procedure, as well as for distinct benchmark datasets that cover both genetic diseases and bio-molecular pathways. The performances for the two prioritization approaches are very similar, although the kernel based algorithm seems to outperform the Order Statistics based algorithm in general. However, it is known the AUC of the cross-validation procedure is likely an over-estimation of the performance for real biological applications; there is therefore the need for different validations.

The OS method was used in simulated predictive studies in order to assess its ability to found back the recently reported disease-gene associations from the literature using genomic data prior to the discovery. The results indicate that the performance for predictive studies is indeed lower than for a cross-validation setup. However, our method can still efficiently rank the novel disease gene (within the top 30% more than 90% of the cases). The OS method was furthermore successfully applied to real biological problems, which proved its usefulness. Through collaboration, it was used to identify a novel DiGeorge syndrome candidate gene YPEL1, to identify 12 novel *atonal* interactors, and to identify a novel CHD gene TAB2. It was furthermore used in 20 external studies to investigate different genetic disorders and has led to important discoveries. Our methods were implemented into publicly available softwares, termed Endeavour (for the Order Statistics based method), and MerKator (for the kernel based method). These tools have been used by other teams worldwide and led to various biological discoveries in human genetics as described above.

There are still several avenues that can be explored in order to enhance the quality of our gene prioritization approaches. At the conceptual level, efforts should be made toward an automatic building of the training set, and through the use of alternative training information. At the algorithmic level, different kernel / network

based approaches can be investigated. At the data level, a method that would prioritize not only genes but also other bio entities such as non coding RNAs, micro RNAs, peptides, chemicals, would represent a very interesting tool for researchers and scientists involved in drug development.

As shown by some the external validations, the integration within sequencing based workflows is also a very interesting and efficient opportunity. Another possibility is the development of a platform for gene prioritization, clustering and classification. In addition, and following our recent experience, the licensing of these tools can be considered. These issues are further discussed in the sections below.

9.1 Conceptual improvements

9.1.1 Training set

The methods we have developed can be classified as novelty detection methods, they differ from classification methods in the absence of a negative training set. A positive training set is however required, its building is a critical step since the quality of the prioritization directly correlates with its quality. In our case, the positive training set simply consists of genes known to be involved in the biological process of interest. There are however several shortcomings. First, assembling a set of genes is a tedious process for a non expert. Several databases that collect disease-gene associations exist but publicly available solutions such as OMIM and GAD are rather incomplete and the building of a training set can not be based only on the consultation of these resources. In addition, literature search engines such as PubMed and GoPubMed have to be used manually to define a more complete training set. In addition, and when available, commercial solutions such as Ingenuity Pathway Analysis (IPA), or Human Gene Mutation Database (HGMD) have to be manually accessed and browsed to enrich these sets.

The development of strategies to help or automatize the construction of the training set is likely to make the overall approach more user-friendly, and therefore more accessible to biologists. It is often easier for the user to define the biological process he is interested in via a list of keywords or even simply its name. A first and easy solution was implemented within Endeavour and consists in retrieving automatically the disease genes from OMIM, but because of its incompleteness and its free-text structure, OMIM can not guarantee the retrieval of all known disease genes. A second option is to use keywords directly to train the algorithm, this technique is mostly used by candidate gene prioritization tools that rely only on text-mining data such as Bitola, PGMapper and GeneProspector. In this setup, the keywords are linked to publications, as opposed to a gene centric approach in which publications are linked to genes. This is off course less suited for other type

of data such as annotation and interaction data for which gene centric approaches have a clear advantage.

These two options can be combined to maximize the quality of the gene set retrieved. Starting from a single keyword, for instance a genomic disease common name, a more complete list of keyword can be obtained through text-mining or through the use of a pre-computed dictionary. A second round of text mining with the enriched list of keywords allows the retrieval of the whole literature that describes the biological process of interest, possibly several hundreds of publications. Interestingly, these documents will also contain the genes that are key for this biological process. In addition, knowledge bases and databases can also be queried to polish the results of the text-mining method. This workflow would automate the creation of the training set, starting from a single keyword and make the process a lot easier for the end user.

9.1.2 Biological entity prioritization

Our prioritization platform is gene centric, meaning that only genes can be prioritized. The system is however using mRNA and protein data that is mapped back to genes, but the data structure is still centered around genes. Moreover, only the protein coding genes can efficiently be prioritized due to a lack of data for the non protein coding genes. However, the cellular mechanisms can not be restricted to the triplet gene-mRNA-protein. The other key players are all the non coding genes and mRNA (*e.g.,* miRNAs), the peptides that are very small amino-acid chains, and various chemical compounds that are also present in the human cells (*e.g.,* drugs). These key players have been ignored so far, but it might be worthwhile including them in the future. The discovery of the importance of miRNAs [182, 201] and of their regulatory function [32, 109, 122, 129] is still recent, but already a lot of data has been collected about them such as where the miRNA genes are located, when they are expressed and which mRNAs they target, and for which purposes. Furthermore, computational tools have been developed to enrich this knowledge with predictions: detect miRNA genes using a cross-species approach or predicting which mRNAs are targeted using sequence similarities. Globally, this represents enough data for our gene prioritization to work accurately on miRNAs. Other bio-entities such as chemical compounds are also of primary interest for example for pharmaceutical and biotech companies that are working in chemistry, chemoinformatics and drug development. These companies have developed or co-developed large databases of chemical compounds and gather data about their function, their effect within cells, and the other cellular players they are interacting with. Once again, this represents a subsequent amount of data that can be used to prioritize chemical compounds. The rationale behind this update is that it is often the case that a researcher is investigating a biological process without knowing exactly beforehand which cellular player he/she is looking for.

9.1.3 Feature selection

Our current framework makes use of kernel based methods to perform gene prioritization. The kernels are calculated beforehand for all the data sources and loaded into memory at run time. This approach reveals to be useful for human and the other species used in that thesis, but can be a limiting factor for species with a very large number of genes (*e.g.*, around 33000 for rice). Loading such large kernels at run time is however very costly when only part of it is needed to solve the prioritization problem. Another option is to calculate the similarities (*e.g.*, using the dot product between the vectors) on the fly to save memory. A clear advantage of this method is that it is possible to include feature selection into the workflow.

In most classification problems, many types of features are first extracted and then fused. Although very effective for classification, kernel based fusion methods do not reduce data dimensionality. There exists techniques to include feature selection within a kernel based framework [124, 225, 13, 147, 173] and they can be investigated in the context of gene prioritization. This feature selection step can be used to tune the kernel towards the problem under investigation, which would not be possible if the kernel has to be computed beforehand.

9.1.4 Kernel fusion scheme

The present dissertation describes the use of Quadratically Constrained Quadratic Program (QCQP) to optimize the weights attributed to each kernel for the fusion. This program leads to very sparse solutions that are almost binary solutions, *i.e.*, one or a few kernels are contributing a lot (high coefficients), while the majority of the kernels are contributing very modestly (low coefficients). We have seen that the main advantage of this approach is its robustness to noise. However, a sparse solution might be suboptimal for real biological problems, for which the user wants each data source to contribute to the problem. A first alternative strategy has been developed, it consists in adding a lower boundary on the weights as an additional constraint. This solution makes the whole approach more sensitive to noise without really reducing the sparsity of the solution. Another alternative strategy, termed L_2-norm MKL, has been developed and is leading to non sparse solutions but the main drawback is then that it is more sensitive to noise.

A future research topic can include the definition of alternative methods to this problem. One possibility is to split the process in two steps, the first step would then be a binary decision about whether the data source should be kept for the second step (informative data source) or whether it should be discarded (noisy or uninformative data source). The second step would consist in a regular optimization between informative data sources, maybe using L_2-norm MKL or other methods.

A second option is the use of different methods for this optimization problem, if the problem can be formulated correctly. The main goal is then to define a method that can discard the uninformative data sources while distributing the weights to avoid the "winner takes all" effect.

9.1.5 Improved statistics

A prioritization run results in a ranking of the candidate genes with the most promising genes at the top. This ranking is provided with a p-value for every candidate gene, this p-value estimates how likely it is to obtain these results by chance alone. However, a problem is that they are candidate set dependent, meaning that p-values tend to be smaller (*i.e.*, more significant) when the size of the candidate gene set increases. This is because the rank ratios are taken into account for the calculation. While this is accounting for the fact that ranking first out of 10 genes is not as good as ranking first out of 1000 genes, this also means that virtually any genome wide prioritization results in highly significant candidate genes whether or not the training was performed with meaningful or randomly selected genes. It is thus difficult to distinguish spurious results from real biologically sounding results. Several prioritization approaches that take this into account have been developed and can be used to redefine our statistics calculation. An example is the bayesian virtual pull-down method developed by Lage *et al.* that is able to discriminate the cases for which no significant results can be obtained at all despite the fact that one candidate gene is still ranked first [128]. One possibility is the extension of our current order statistics approach to account for the number of candidate genes. Currently the single parameter is the number of ranks ratios combined, and based on its value, a gamma or a beta distribution with different parameters are used to derive the final p-values. Another possibility is to use randomization using the same candidate set to correct the p-values for its size, this will however require more computing power since each randomization means another prioritization.

9.2 Technical improvements

9.2.1 Simpler inputs

The inputs of our methods are a set of training genes, a set of candidate genes (the while genome can be used if no region can be defined), and the data sources to use in the prioritization process. The selection of the data sources not only requires knowledge about the content of these data sources but also knowledge about which ones should be used in order to obtain meaningful results. This is something that is difficult to estimate beforehand, however one can determine the optimal set of

data sources through leave-one-out cross-validation. Those would be the ones for which a good performance, *i.e.*, AUC, is observed. The proposed strategy is that, first, a cross-validation on the training set is run, second, the optimal set of data sources is defined and, third, the regular prioritization takes place. In addition, this cross-validation would also allow the detection of outliers within the training set, these would be genes that do not seem to belong to the same biological process and therefore are only adding noise to the predictions. Of interest, the removal of these outliers and the automatic selection of the data sources to use is likely to have a positive impact on the global performance. To conclude, in the future, the user would be able to input only the two gene sets without having to consider which genomic data source should be used. An additional parameter would still be the use of the cross-species version in which data from model organism is used in conjunction with human data.

9.2.2 Detailed results

The end user of computational tools such as Endeavour are bioinformaticians, biologists and geneticist, and we aim at integrating our method within their workflows. This means that not only it should be easy to use and to integrate but also that it requires simple inputs and provide detailed output. The detailed outputs is the description of which genes are considered as promising and, more importantly, why. It is indeed crucial for the user to understand and interpret the results returned to him/her. With the current version of Endeavour, only limited information is returned to the user: the global ranking of the candidate genes, plus an additional ranking per data source used in the prioritization. These additional rankings are already providing information about which data sources are contributing to the overall ranking but this is not sufficient. What users would like to see is information for each data source about why a gene is ranked that way: what is the underlying genomic data, and what is linking that candidate to the known genes used for training.

In the future, the Endeavour web based interface should be enhanced with new modules that display such information to the user. One characteristic of our kernel based algorithm is the uneasy interpretability of the results. Indeed, turning genomic data into kernel matrices means that only the distance between the genes are kept (and therefore their similarities), while the underlying genomic data is lost. Furthermore, the use of the SVM algorithm means that it is not possible to establish direct relationship between the candidate genes and the training genes. So it means that alternative strategies will have to be developed. One possibility is to use a network representation of both the training genes and the candidate genes, so that relations can easily be identified. This network can then be enriched with additional information stating which genomic data underlie these relations. This is similar to what is done in String that combines multiple data sources together in

a network representation. The main difference is that, in our case, only human data should be included in the network and no prediction should be made (String includes prediction based on model organisms), furthermore, String mostly relies on Text mining data, we would like to take more data sources into consideration. This strategy has the advantage that the kernel and the network representations are very close and that there exist methods to go from one to the other.

9.2.3 Extension

Several possibilities exist to extend our method. A first option is to add more genomic data sources such as chemical data and phenotypic data. The inclusion of more data sources, that are possibly not covered by current data sources can strengthen the approach. For instance, the inclusion of miRNA specific data would allow predictions based on miRNA based regulatory mechanisms that are distinct from the mechanisms described by our TFBS database. It would also worth to complete even further the description of the regulatory mechanisms by including epigenetic data such as DNA methylation and chromatin remodeling. In addition, the inclusion of more phenotypic data would probably be beneficial for the prioritization of disease candidate genes, the use of chemical data is also very likely to enhance the performance for disease studies since most of this data comes from drug development research. Some data sources are very small and therefore can not directly be used as such but need to be merged with other sources before. For example, there exist many protein-protein interaction databases that are disease specific (*e.g.*, AlzGene [33]), or that focus on a subset of the PPIs (*e.g.*, MIPS focuses on physical interactions from the scientific literature). These datasets are sometimes capturing a unique biological signal that is not covered by other more generalist databases. The integration of small datasets into a larger global dataset can be realized using known integration algorithms [246, 134] or by developing novel integration schemes. As long as the data can be summarized into an existing format (*e.g.*, vector based, annotation based, network based), and transformed into kernels, the integration should be easy to realize and should enrich the database underlying Endeavour.

Another option is to include more organisms for which enough genomic data can be gathered. Our goal is to combine together species that are closely related so that prediction in one species can be made by using data from the other species. Unicellular eukaryotic organisms, such as yeast, have been studied a lot because the yeast cells are easy to grow and because the yeast genome was early sequenced [52]. There is therefore a lot of data available including whole genome knockout experiments that are almost unique for eukaryotes. However, yeast and human are too far away in the phylogenetic tree, thus including yeast data is unlikely to help a lot. Most of the model organisms have already been integrated, however, primates and mammalian species also represent possible extensions (*e.g.*, chimp).

This solution is very attractive: the cost is relatively reasonable if we assume that the same algorithms can be used. Then, data sources need to be collected only once for the novel species, and the same update system can be reused to update the data regularly. The gain can however be very significant, adding more species allows researchers who are working with these species to prioritize candidate genes but also allows human geneticists to get more accurate results though cross-species transfer.

9.2.4 Several objectives, a single platform

In the recent years, our bioinformatics group has gained expertise in gene prioritization, in gene clustering and in gene classification. However, most of this work was performed by different people on different data and with different aims, although the underlying methods are very similar (most of the time, kernel based methods are used). In addition, most of methods implemented in this work have not been turned into publicly available tools. One objective is to build a common platform for gene prioritization, clustering and classification. The core of this platform would be based on kernel methods that have proved to be useful for this tasks [34, 270, 269, 58]. By gathering together different algorithms that can perform different computational tasks on the same data, we can greatly enhance the number of potential users and build a reference platform in computational biology.

9.3 More applications

The application to real biological problems described in this thesis not only proved that our approach conceptually sounds but also that its integration within wet-lab workflows is doable and beneficial. One of the future objectives is to find novel biological applications. To this end, we can rely on already existing scientific collaborations with the following groups from K.U.Leuven:

1. Laboratory of Computational Biology, headed by Prof. Aerts
2. Laboratory for Cytogenetics and Genome Research, headed by Prof. Vermeesch
3. Laboratory of Neurogenetics, headed by Prof. Hassan
4. Laboratory for Genetics of Human Development, headed by Prof. Devriendt

But of course, by making our method publicly available, and easily integrable, we hope that other research teams will also use the tool and report discoveries.

To ease that process, we have already described how Endeavour can be used in conjunction with array CGH data [4, 232], or with a forward genetic screen [7]. This can further be extended with its inclusion into a sequencing based workflow.

9.3.1 A sequencing based workflow

The recent development of the new generation sequencing technology opens the door for a new sequencing era. This faster technology available for a reduced cost will soon allow the sequencing of hundreds of genomes in no time. A corollary is that a massive amount of data will be generated, and that computational tools are needed in order to analyze, and organize this data.

Beside the softwares needed to preprocess and analyze the raw data, prioritization tools might still be needed to identify the most interesting SNPs hidden among the data. One possibility to identify disease causing genes using new generation sequencing technology is to sequence the genome of patients and to analyze the SNPs that are observed. Preliminary studies have shown that thousands or even millions of SNPs are identified as potentially linked to the disease under study [146]. There is thus a need to prioritize them in order to find the most promising SNPs that are likely to be associated to the disease. SNPs prioritization very much resemble gene prioritization except that additional SNP specific information can be used such as the change in protein structure that the SNPs are causing, or whether or not it is within a known functional domain. Several SNP prioritization strategies have been developed in parallel of the gene prioritization methods, but so far only a limited number of these can prioritize both genes and SNPs (*e.g.*, SNPs3D [273]) but are still performing the two in an independent manner. The proposed strategy is to tunnel both methods to prioritize genes first and SNPs in a second place.

9.4 Long term objectives

This section describes a long term objective, the creation of a computational platform dedicated to gene prioritization, clustering and classification based on our existing prioritization tool Endeavour. By implementing the changes described in this chapter, we aim at developing a product that we can license to pharmaceutical and biotech companies and to genetics laboratories. Our two first licensing opportunities have taught us that there is an interest in our product although we are not yet responding exactly to the current market's needs. The responses furthermore indicate that the emerging market is not yet sufficiently mature to support a full scale commercial activity. However, we expect the demand to increase in the next five years so that our product will be commercially viable in the mid-term. In addition to the licensing, we also want to contract the corresponding

services such as software installation, user formation and counseling. The following subsections describe our past licensing opportunities, and based on that, our projected business plan and the IP situation. This could serve as a preliminary draft to define a more complete business plan.

9.4.1 Licensing opportunities

Our group has experience in the exploitation of research since we already went twice through the exercise of licensing our software Endeavour. The first licensee was a major pharmaceutical company, Novartis (LRD license number 2008/1397, starts on 20/10/2008, 3 months evaluation, 10k€). The agreement with Novartis stated that they would use and evaluate the software within a period of three months that could eventually be followed by a full license (one year renewable). The trial period took place between October 2008 and February 2009 and unfortunately did not lead to a full license agreement. They mention that the approach is scientifically sounding and that the results of the validation were very interesting but also raised a number of concerns that made us think that our product was not yet completely ready. This showed that there is still a mismatch between what we can deliver and what the market is willing to pay for.

CropDesign from BASF plant has also signed a license more recently (LRD license number 2010/135, starts on 15/01/2010, 3 months evaluation, 10k€ support and 3k€ license fee). Similarly to the first case, the license includes a trial period of three months needed to evaluate the software. A difference with our first experience is that the license includes a contract that covers the services that we offer in parallel. These services are the on site installation of the software, the formation of the end users, and the consulting for the elaboration of the validation process.

9.4.2 Business plan

At this early stage, several business models are under study. The first one is the creation of a bioinformatics spin-off that will license the product, and the associated services. A more ambitious idea is the inclusion of our product into a more high level workflow and therefore the creation of a biotech spin-off. In any case, the creation of a spin-off should only be investigated when a precise market has been defined and when the product is no longer in development. This is why a second option have been defined and, in fact, represents what has been done so far: licensing through the university, via the LRD department. With this option, the services can also be contracted so that there are no differences with the first option from the licensing/contracting point of view. This second option has the advantage of reducing the risk linked to the spin-off creation. The last option also shares this advantage, it is a collaboration with an industrial partner through an O&O project

to co-develop and co-license the product together. This option is similar to the second one, except that the partner is part of the industry and potentially already owns a product and knows its associated market. All these options are considered now that the project is still in a early phase but future will tell us which one is more appropriate.

9.4.3 Intellectual property

Our main objective is the development of a computational platform, that is a software suite. Patenting softwares is not highly effective for several reasons:

1. The copyright that applies for the code underlying the software already represents an efficient protection.
2. These patents can often be circumvented through slight modifications of the embodiment.
3. Most of the companies that develop softwares have adopted the speed to market paradigm, in order to cut lead time and to stay innovative. In this model, patenting the software is not always a suitable strategy.

As for the IP, there exist commercially available computational tools that represent potential competitors. A first category contains the microarray analysis tools (*e.g.*, GeneSpring, ArrayAssist®, ArrayStar, Mapix, Qlucore Omics Explorer, Axon GenePix, Spotfire® and Pathway Architect™). These tools are similar to our platform because they allow users to analyze large list of genes. There are however several differences:

1. They are not making use of various genomic data and usually rely on expression data alone (or in combination with phenotypic data). Our approach is to combine many data sources, we also include, for instance, literature data, functional annotations, sequences and regulatory information.
2. We want to use advanced machine learning that have been developed recently in academia and that have not yet been implemented in microarray analysis softwares.
3. Microarray analysis is often reduced to clustering/classification of the genes/conditions. In addition, we propose prioritization the candidate genes with respect to one biological process of interest.

A second category contains the biological knowledge bases such as Ingenuity Pathways Analysis and GeneGo. These databases are very useful since they contain

high quality genomic data which is in most of the cases manually curated by experts in the field. Their main drawback is that they represent passive knowledge bases. We believe that our approach will add significant value to this field since the knowledge base is in our case a basis to infer new associations and to make predictions.

To conclude, this thesis describes the development and the validation of a gene prioritization method, but many different avenues still have to be explored in order to enhance its ability and applicability, to prove its usefulness, and eventually to commercialize it.

Appendix A

Algorithm behind Endeavour

This appendix presents in details the Endeavour algorithm. It complements figure 1.4. The algorithm comprises three steps: training, scoring, and fusion.

A.1 Training

For the training step, the aim is to modelize the process of interest, more precisely through the genes that are known to play a role in this process. Several data type are defined and each one has its own modeling method.

A.1.1 Annotation data

For annotation data, the model is the set of features (*i.e.*, annotation terms) that are over-represented in the training set when compared to the genome. The over-representation is calculated using the binomial distribution as an approximation of the quadratic distribution. More precisely, the model contains the over-represented terms together with p-values that correspond to the probability to observe such over-representation by chance alone. The p-values are corrected for multiple testing with the Bonferroni correction.

A.1.2 Vector based data

For vector based data, the vectors of the training genes are retrieved and an average vector is calculated. This vector represents the model. An exception is

made for expression data, for which calculating an average vector is not leading to a good model because in most of the cases, the training genes are not perfectly co-expressed. Therefore, for expression data, the model is simply the collection of the expression profiles of the training genes.

A.1.3 Interaction data

For interaction data, the training genes and their interacting partners are collected from the global PPI network. This subnetwork represents then the model.

A.1.4 Sequence data

For sequence data, the sequence similarities between all genes pairwise is calculated beforehand. The model consists in extracting from this huge matrix the submatrix that contains the hits between the training genes and any other gene.

A.1.5 Precomputed data

There is no need to train for precomputed data.

A.1.6 Special cases

For literature data, the text mining procedure is performed beforehand to connect each gene with a number of keywords. Although, it is a bit similar to the annotation data sources, there are differences. The main one is that the gene-term associations are not binary but are provided with a score (TFIDF) that represents the quality of the over-representation of the keyword in the publications linked to the genes when compared to background publications. So, the vector based modeling method is used, that is an average vector is created.

A.2 Scoring

The scoring step consists in assigning, to each candidate gene, a score that reflects its similarity to the model built in the first step. The candidate genes are scored independently of each other and then ranked according to their scores so that the most promising genes are always on top.

A.2.1 Annotation data

For annotation data, the p-values of the candidate gene annotation terms that are also present in the model are combined using Fisher's omnibus. Using χ^2, a global p-value is derived from the Fisher's omnibus score. This p-value is used as a score for the candidate gene. A small p-value means that the candidate gene is associated with many of the terms that are in the model, and therefore that the candidate gene is a promising gene.

A.2.2 Vector based data

For vector based data, a very simple approach is used: the cosine of the angle or the Pearson correlation between the candidate gene profile and the model profile is calculated and used as a score for the candidate. A small cosine or a high correlation means that the candidate gene profile is highly similar to the model and therefore that the candidate gene is a promising gene. An exception is made for expression data, for which the model contains as many profiles as they are training genes. In this case, the candidate gene profile is compared to each training profile (using again the cosine of the angle or the Pearson correlation), then only the best 50% scores as kept (*e.g.*, for six training profiles, only the best three cosines are kept for each candidate gene). These scores are then combined simply by averaging them to obtain the final score. A small final score means that the candidate gene profile is similar to half of the training profiles, meaning that it is indeed an interesting candidate gene.

A.2.3 Interaction data

For interaction data, the candidate gene and its interacting partners in the global PPI network are retrieved. This subnetwork is then compare with the model subnetwork. In particular, the overlap between the two is considered and the score is the size of this overlapping region divided by the size of the candidate gene subnetwork (to correct for genes that have many interacting partners). A high score means that most of the candidate gene subnetwork elements are also found in the model subnetwork and therefore that the candidate gene is a promising gene.

A.2.4 Sequence data

For sequence based data, the sequence alignment scores (from Blast) between the training genes and the candidate genes are retrieved. For each candidate, only the

best similarity is kept as the final score. So the most interesting candidate genes are the ones that have high sequence similarity with one of the training gene.

A.2.5 Precomputed data

For precomputed data, the score has already been computed and can directly be retrieved to rank the candidate genes. Notice that the training genes have no influence on these scores (as opposed to sequence data for instance).

A.3 Data fusion

The data fusion process is performed using the Order Statistics (OS). For each candidate gene, the ranks obtained are transformed into rank ratios by dividing the rank with the number of genes ranked for this data source. Due to the missing values, the number of genes ranked varies among the data sources, and by calculating the rank ratios, we take this into account. The rank ratios are then sorted from the smallest to the largest and then combined into a single value using the OS. An alternative formula is used to fasten the calculation, defined in Aerts *et al.* [4]. St this stage, each candidate gene is associated with a score but these scores can not be compared directly they are not all derived from the same number of rank ratios (again due to the missing values). The next step is to derive p-values from these scores by using Beta and Gamma distributions whose parameters depend on the number of rank ratios. The parameters have been estimated beforehand by looking at the distributions of the scores of thousands of prioritizations. For a single candidate gene, a final p-value can be derived from the rank ratios by estimating the probability to obtain these rank ratios by chance alone (using the Beta/Gamma distribution). The final ranking is made according to these p-values.

Appendix B

Lists of candidate genes

Candidate genes				
Rab18	Rab5a	Rab22a	Kras	Bet1
Arf5	Stx7	Pacsin2	Rhog	Sec24c
Rab4b	Vamp8	Stx12	Akt2	Dnm2
Vamp3	Rab14	Golga3	Rab3d	Vti1b
Cdc42	Rab6	Npepl1	Dbnl	Snapap
Arf4	Sybl1	Scfd1	Rab10	
Rab3a	Snap29	Stx6	Rab6b	
Rab1	Vps45	Rab5c	Eea1	
Rab2	Rab8a	Dock1	Ykt6	

Table B.1: The 41 candidate genes obtained through gene prioritization by Adachi *et al.* who studied the adipocyte mediated energy metabolism [1].

Candidate genes				
ABCC9	GRIA4	KCNJ3	KCNK6	NSF
ATP1A2	GRIK2	KCNJ5	KCNN2	PRKCABP
ATP1A4	GRIK4	KCNJ6	KCNN4	PRKCG
ATP1B2	GRIN1	KCNJ8	KCNV2	PSD95
GIRK3	GSR	KCNK16	KCTD13	RIPK1
GLUR6	KCNJ10	KCNK17	KCTD17	SLC24A3
GRIA2	KCNJ16	KCNK5	KCTD3	

Table B.2: The 34 candidate genes obtained through gene prioritization by Poot *et al.* who identify recurrent Copy Number Changes (CNCs) in mentally retarded children [191].

Candidate genes				
NPY1R	CTSO	ESR1	SCARB1	PPGB
NPY2R	GLRB	ENPP1	TCF1	HNF4A
NPY5R	TLR2	FLI1	NCOR2	PTGIS
CPE	GRM1	KCNJ5	GNAS	
FGB	OPRM1	ROBO4	LAMA5	
FGG	LATS1	AACS	PCK1	

Table B.3: The 27 candidate genes obtained through gene prioritization by Elbers *et al.* who study type 2 diabetes and obesity [73].

Candidate genes				
COL1A1	KRT10	SEMA6A	NDRG3	HCRT
LBP	FKBP10	EPB41L1	GHRH	KRT14
CCR7	ATP6V0A1	NNAT	SCAND1	KRT35
HSD17B4	IGFBP4	CTNNBL1	KRT16	NAGLU
KRT13	KRT17	TGIF2	ACLY	ARNTL2
KRT15	FMR1	RBL1	KCNH4	MED21
KRT19	USP9X	SLA2	CNTNAP1	

Table B.4: The 34 candidate genes obtained through gene prioritization by Liu *et al.* who propose new candidate for areal BMD (aBMD) and areal bone size (ABS) that are both risk factors for osteoporosis [139].

Bibliography

[1] Jun Adachi, Chanchal Kumar, Yanling Zhang, and Matthias Mann. In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Molecular & Cellular Proteomics: MCP*, 6(7):1257–1273, July 2007. PMID: 17409382.

[2] E A Adie, R R Adams, K L Evans, D J Porteous, and B S Pickard. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics (Oxford, England)*, 22(6):773–774, March 2006. PMID: 16423925.

[3] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55, 2005. PMID: 15766383.

[4] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Léon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, Peter Carmeliet, and Yves Moreau. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544, May 2006. PMID: 16680138.

[5] Stein Aerts, Peter Van Loo, Gert Thijs, Herbert Mayer, Rainer de Martin, Yves Moreau, and Bart De Moor. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Research*, 33(Web Server issue):W393–396, July 2005. PMID: 15980497.

[6] Stein Aerts, Gert Thijs, Bert Coessens, Mik Staes, Yves Moreau, and Bart De Moor. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Research*, 31(6):1753–1764, March 2003. PMID: 12626717.

[7] Stein Aerts, Sven Vilain, Shu Hu, Leon-Charles Tranchevent, Roland Barriot, Jiekun Yan, Yves Moreau, Bassem A Hassan, and Xiao-Jiang Quan. Integrating computational biology and forward genetics in drosophila. *PLoS Genetics*, 5(1):e1000351, January 2009. PMID: 19165344.

[8] Sumeet Agarwal, Charlotte M Deane, Mason A Porter, and Nick S Jones. Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Computational Biology*, 6(6):e1000817, 2010. PMID: 20585543.

[9] Andrey Alexeyenko and Erik L L Sonnhammer. Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Research*, 19(6):1107–1116, June 2009. PMID: 19246318.

[10] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L Papadopoulos, Martin Reczko, and Artemis G Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics (Oxford, England)*, 25(23):3049–3055, December 2009. PMID: 19789267.

[11] I Altintas, C Berkley, E Jaeger, M Jones, B Ludascher, and S Mock. Kepler: an extensible system for design and execution of scientific workflows. In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*, pages 423–424, 2004.

[12] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990. PMID: 2231712.

[13] Carlos Alzate and Johan A K Suykens. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):335–347, February 2010. PMID: 20075462.

[14] Lars Andersson, Greta Petersen, Per Johnson, and Fredrik Ståhl. A web tool for finding gene candidates associated with experimentally induced arthritis in the rat. *Arthritis Research & Therapy*, 7(3):R485–492, 2005. PMID: 15899035.

[15] A Appel, A L Horwitz, and A Dorfman. Cell-free synthesis of hyaluronic acid in marfan syndrome. *The Journal of Biological Chemistry*, 254(23):12199–12203, December 1979. PMID: 500705.

[16] Roland Arnold, Thomas Rattei, Patrick Tischler, Minh-Duc Truong, Volker Stümpflen, and Werner Mewes. SIMAP–the similarity matrix of proteins. *Bioinformatics (Oxford, England)*, 21 Suppl 2:ii42–46, September 2005. PMID: 16204123.

[17] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, May 2000. PMID: 10802651.

[18] N Azpiazu, P A Lawrence, J P Vincent, and M Frasch. Segmentation and specification of the drosophila mesoderm. *Genes & Development*, 10(24):3183–3194, December 1996. PMID: 8985186.

[19] Francisco Azuaje, Yvan Devaux, and Daniel Wagner. Computational biology for cardiovascular biomarker discovery. *Briefings in Bioinformatics*, 10(4):367–377, July 2009. PMID: 19276200.

[20] Francisco Azuaje, Yvan Devaux, and Daniel R Wagner. Coordinated modular functionality and prognostic potential of a heart failure biomarker-driven interaction network. *BMC Systems Biology*, 4:60, 2010. PMID: 20462429.

[21] Francisco Azuaje, Yvan Devaux, and Daniel R Wagner. Integrative pathway-centric modeling of ventricular dysfunction after myocardial infarction. *PloS One*, 5(3):e9661, 2010. PMID: 20300185.

[22] G D Bader, I Donaldson, C Wolting, B F Ouellette, T Pawson, and C W Hogue. BIND–The biomolecular interaction network database. *Nucleic Acids Research*, 29(1):242–245, January 2001. PMID: 11125103.

[23] Gary D Bader, Doron Betel, and Christopher W V Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, January 2003. PMID: 12519993.

[24] R L Baehner, L M Kunkel, A P Monaco, J L Haines, P M Conneally, C Palmer, N Heerema, and S H Orkin. DNA linkage analysis of x chromosome-linked chronic granulomatous disease. *Proceedings of the National Academy of Sciences of the United States of America*, 83(10):3398–3401, May 1986. PMID: 3010296.

[25] A E Bale and K P Yu. The hedgehog pathway and basal cell carcinomas. *Human Molecular Genetics*, 10(7):757–762, April 2001. PMID: 11257109.

[26] Giuseppe Barbesino, Yaron Tomer, Erlinda S. Concepcion, Terry F. Davies, and David A. Greenberg. Linkage analysis of candidate genes in autoimmune thyroid disease. II. selected Gender-Related genes and the X-Chromosome. *J Clin Endocrinol Metab*, 83(9):3290–3295, September 1998.

[27] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, and Ron Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research*, 37(Database issue):D885–890, January 2009. PMID: 18940857.

[28] Roland Barriot, Jeroen Breckpot, Bernard Thienpont, Sylvain Brohée, Steven Van Vooren, Bert Coessens, Léon-Charles Tranchevent, Peter Van

Loo, Marc Gewillig, Koenraad Devriendt, and Yves Moreau. Collaboratively charting the gene-to-phenotype network of human congenital heart defects. *Genome Medicine*, 2(3):16, March 2010. PMID: 20193066.

[29] John E Beaver, Murat Tasan, Frank Gibbons, Weidong Tian, Timothy R Hughes, and Frederick P Roth. FuncBase: a resource for quantitative gene function annotation. *Bioinformatics (Oxford, England)*, May 2010. PMID: 20495000.

[30] Isabelle Vanden Bempt, Maria Drijkoningen, and Christiane De Wolf-Peeters. The complexity of genotypic alterations underlying HER2-positive breast cancer: an explanation for its clinical heterogeneity. *Current Opinion in Oncology*, 19(6):552–557, November 2007. PMID: 17906451.

[31] Asa Ben-Hur and William Stafford Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics (Oxford, England)*, 21 Suppl 1:i38–46, June 2005. PMID: 15961482.

[32] E Bernstein, A A Caudy, S M Hammond, and G J Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366, January 2001. PMID: 11201747.

[33] Lars Bertram, Matthew B McQueen, Kristina Mullin, Deborah Blacker, and Rudolph E Tanzi. Systematic meta-analyses of alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics*, 39(1):17–23, January 2007. PMID: 17192785.

[34] Tijl De Bie, Léon-Charles Tranchevent, Liesbeth M M van Oeffelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics (Oxford, England)*, 23(13):i125–132, July 2007. PMID: 17646288.

[35] Daniel Blankenberg, James Taylor, Ian Schenck, Jianbin He, Yi Zhang, Matthew Ghent, Narayanan Veeraraghavan, Istvan Albert, Webb Miller, Kateryna D Makova, Ross C Hardison, and Anton Nekrutenko. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Research*, 17(6):960–964, June 2007. PMID: 17568012.

[36] Joseph Bockhorst and Mark Craven. Markov networks for detecting overlapping elements in sequence data. *Neural Information Processing Systems*, 2005.

[37] R J Boucek, N L Noble, Z Gunja-Smith, and W T Butler. The marfan syndrome: a deficiency in chemically stable collagen cross-links. *The New England Journal of Medicine*, 305(17):988–991, October 1981. PMID: 7278923.

[38] Pascal Braun, Edward Rietman, and Marc Vidal. Networking metabolites and diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29):9849–9850, July 2008. PMID: 18632571.

[39] Terry A Braun, Suma P Shankar, Steve Davis, Brian O'Leary, Todd E Scheetz, Abbot F Clark, Val C Sheffield, Thomas L Casavant, and Edwin M Stone. Prioritizing regions of candidate genes for efficient mutation screening. *Human Mutation*, 27(2):195–200, February 2006. PMID: 16395665.

[40] Frank C Brosius, Charles E Alpers, Erwin P Bottinger, Matthew D Breyer, Thomas M Coffman, Susan B Gurley, Raymond C Harris, Masao Kakoki, Matthias Kretzler, Edward H Leiter, Moshe Levi, Richard A McIndoe, Kumar Sharma, Oliver Smithies, Katalin Susztak, Nobuyuki Takahashi, and Takamune Takahashi. Mouse models of diabetic nephropathy. *Journal of the American Society of Nephrology: JASN*, 20(12):2503–2512, December 2009. PMID: 19729434.

[41] Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artif. Intell. Med.*, 33(2):139–155, 2005.

[42] J Burn, P Brennan, J Little, S Holloway, R Coffey, J Somerville, N R Dennis, L Allan, R Arnold, J E Deanfield, M Godman, A Houston, B Keeton, C Oakley, O Scott, E Silove, J Wilkinson, M Pembrey, and A S Hunter. Recurrence risks in offspring of adults with major heart defects: results from first cohort of british collaborative study. *Lancet*, 351(9099):311–316, January 1998. PMID: 9652610.

[43] William S Bush, Scott M Dudek, and Marylyn D Ritchie. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 368–379, 2009. PMID: 19209715.

[44] P H Byers, R C Siegel, K E Peterson, D W Rowe, K A Holbrook, L T Smith, Y H Chang, and J C Fu. Marfan syndrome: abnormal alpha 2 chain in type i collagen. *Proceedings of the National Academy of Sciences of the United States of America*, 78(12):7745–7749, December 1981. PMID: 6950413.

[45] Giulio Calcagni, M Cristina Digilio, Anna Sarkozy, Bruno Dallapiccola, and Bruno Marino. Familial recurrence of congenital heart disease: an overview and review of the literature. *European Journal of Pediatrics*, 166(2):111–116, February 2007. PMID: 17091259.

[46] Borja Calvo, Núria López-Bigas, Simon J Furney, Pedro Larrañaga, and Jose A Lozano. A partially supervised classification approach to dominant and

recessive human disease gene prediction. *Computer Methods and Programs in Biomedicine*, 85(3):229–237, March 2007. PMID: 17258838.

[47] Sarah Calvo, Mohit Jain, Xiaohui Xie, Sunil A Sheth, Betty Chang, Olga A Goldberger, Antonella Spinazzola, Massimo Zeviani, Steven A Carr, and Vamsi K Mootha. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nature Genetics*, 38(5):576–582, May 2006. PMID: 16582907.

[48] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science (New York, N.Y.)*, 321(5886):263–266, July 2008. PMID: 18621671.

[49] Antonios Chatzigeorgiou, Antonios Halapas, Konstantinos Kalafatakis, and Elli Kamper. The use of animal models in the study of diabetes mellitus. *In Vivo (Athens, Greece)*, 23(2):245–258, April 2009. PMID: 19414410.

[50] Jing Chen, Huan Xu, Bruce J Aronow, and Anil G Jegga. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, 8:392, 2007. PMID: 17939863.

[51] Dean Cheng, Craig Knox, Nelson Young, Paul Stothard, Sambasivarao Damaraju, and David S Wishart. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, 36(Web Server issue):W399–405, July 2008. PMID: 18487273.

[52] J M Cherry, C Ball, S Weng, G Juvik, R Schmidt, C Adler, B Dunn, S Dwight, L Riles, R K Mortimer, and D Botstein. Genetic and physical maps of saccharomyces cerevisiae. *Nature*, 387(6632 Suppl):67–73, May 1997. PMID: 9169866.

[53] Ching-Lung Cheung, Pak C Sham, Vivian Chan, Andrew D Paterson, Keith D K Luk, and Annie W C Kung. Identification of LTBP2 on chromosome 14q as a novel candidate gene for bone mineral density variation and fracture risk association. *The Journal of Clinical Endocrinology and Metabolism*, 93(11):4448–4455, November 2008. PMID: 18697872.

[54] Bert Coessens. *Data integration techniques for molecular biology research.* PhD thesis, Katholieke Universiteit Leuven – Faculteit Ingenieurswetenschappen Arenbergkasteel, B-3001 Heverlee (Belgium), 2006.

[55] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, October 2004. PMID: 15496913.

[56] C Correns. Untersuchungen über die xenien bei zea mays. *Berichte der Deutsche Botanische Gesellschaft*, 17:410–418, 1899.

[57] An Crepel, Jean Steyaert, Wouter De la Marche, Veerle De Wolf, Jean-Pierre Fryns, Ilse Noens, Koen Devriendt, and Hilde Peeters. Narrowing the critical deletion region for autism spectrum disorders on 16p11.2. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, January 2011. PMID: 21225995.

[58] Anneleen Daemen, Olivier Gevaert, Fabian Ojeda, Annelies Debucquoy, Johan Ak Suykens, Christine Sempoux, Jean-Pascal Machiels, Karin Haustermans, and Bart De Moor. A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, 1(4):39, 2009. PMID: 19356222.

[59] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, Pittsburgh, Pennsylvania, 2006. ACM.

[60] H De Vries. Sur la loie de disjonction des hybrides. *Comptes Rendue Hebdemodaires, Acad. Sci. Paris*, 130:845–847, 1900.

[61] Matthieu Defrance and Hélène Touzet. Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC Bioinformatics*, 7:396, 2006. PMID: 16945132.

[62] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5):P3, 2003. PMID: 12734009.

[63] Katrijn Van Deun, Age K Smilde, Mariët J van der Werf, Henk A L Kiers, and Iven Van Mechelen. A structured overview of simultaneous component based data integration. *BMC Bioinformatics*, 10:246, 2009. PMID: 19671149.

[64] Thomas Dhollander, Qizheng Sheng, Karen Lemmens, Bart De Moor, Kathleen Marchal, and Yves Moreau. Query-driven module discovery in microarray data. *Bioinformatics (Oxford, England)*, 23(19):2573–2580, October 2007. PMID: 17686800.

[65] Lori E Dodd and Margaret S Pepe. Partial AUC estimation and regression. *Biometrics*, 59(3):614–623, September 2003. PMID: 14601762.

[66] Andreas Doms and Michael Schroeder. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Research*, 33(Web Server issue):W783–786, July 2005. PMID: 15980585.

[67] Chris Drummond and Robert C. Holte. Explicitly representing expected cost: an alternative to ROC representation. In *Proceedings of the sixth ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 198–207, Boston, Massachusetts, United States, 2000. ACM.

[68] Chris Drummond and Robert C. Holte. What roc curves can't do (and cost curves can). *Proceedings of the ROC Analysis in Artificial Intelligence*, 2004.

[69] Valérie Dupé, Lucie Rochard, Sandra Mercier, Yann Le Pétillon, Isabelle Gicquel, Claude Bendavid, Georges Bourrouillou, Usha Kini, Christel Thauvin-Robinet, Timothy P. Bohan, Sylvie Odent, Christèle Dubourg, and Véronique David. NOTCH, a new signaling pathway implicated in holoprosencephaly. *Human Molecular Genetics*, 20(6):1122 –1131, March 2011.

[70] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, January 2002. PMID: 11752295.

[71] Karen Eilbeck and Suzanna E. Lewis. Sequence ontology annotation guide. *Comparative and Functional Genomics*, 5(8):642–647, December 2004. PMID: 18629179 PMCID: 2447471.

[72] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44, 2005. PMID: 15892872.

[73] Clara C Elbers, N Charlotte Onland-Moret, Lude Franke, Anne G Niehoff, Yvonne T van der Schouw, and Cisca Wijmenga. A strategy to search for common obesity and type 2 diabetes genes. *Trends in Endocrinology and Metabolism: TEM*, 18(1):19–26, February 2007. PMID: 17126559.

[74] Roberta Epis, Fabrizio Gardoni, Elena Marcello, Armando Genazzani, Pier Luigi Canonico, and Monica Di Luca. Searching for new animal models of alzheimer's disease. *European Journal of Pharmacology*, 626(1):57–63, January 2010. PMID: 19836370.

[75] F Erdogan, L A Larsen, L Zhang, Z Tümer, N Tommerup, W Chen, J R Jacobsen, M Schubert, J Jurkatis, A Tzschach, H-H Ropers, and R Ullmann. High frequency of submicroscopic genomic aberrations detected by tiling path array comparative genome hybridisation in patients with isolated congenital heart disease. *Journal of Medical Genetics*, 45(11):704–709, November 2008. PMID: 18713793.

[76] Helen V Firth, Shola M Richards, A Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M Pettett, and Nigel P Carter. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *American Journal of Human Genetics*, 84(4):524–533, April 2009. PMID: 19344873.

[77] Lude Franke, Harm van Bakel, Like Fokkens, Edwin D de Jong, Michael Egmont-Petersen, and Cisca Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American Journal of Human Genetics*, 78(6):1011–1025, June 2006. PMID: 16685651.

[78] Tancred Frickey and Georg Weiller. Mclip: motif detection based on cliques of gapped local profile-to-profile alignments. *Bioinformatics (Oxford, England)*, 23(4):502–503, February 2007. PMID: 17127680.

[79] T K B Gandhi, Jun Zhong, Suresh Mathivanan, L Karthick, K N Chandrika, S Sujatha Mohan, Salil Sharma, Stefan Pinkert, Shilpa Nagaraju, Balamurugan Periaswamy, Goparani Mishra, Kannabiran Nandakumar, Beiyi Shen, Nandan Deshpande, Rashmi Nayak, Malabika Sarker, Jef D Boeke, Giovanni Parmigiani, Jörg Schultz, Joel S Bader, and Akhilesh Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293, March 2006. PMID: 16501559.

[80] Kyle J Gaulton, Karen L Mohlke, and Todd J Vision. A computational system to select candidate genes for complex human traits. *Bioinformatics (Oxford, England)*, 23(9):1132–1140, May 2007. PMID: 17237041.

[81] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004. PMID: 15461798.

[82] Richard A George, Jason Y Liu, Lina L Feng, Robert J Bryson-Richardson, Diane Fatkin, and Merridee A Wouters. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Research*, 34(19):e130, 2006. PMID: 17020920.

[83] Olivier Gevaert. *A Bayesian network integration framework for modeling biomedical data*. PhD thesis, Katholieke Universiteit Leuven – Faculteit Ingenieurswetenschappen Arenbergkasteel, B-3001 Heverlee (Belgium), 2008.

[84] Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics (Oxford, England)*, 22(14):e184–190, July 2006. PMID: 16873470.

[85] Enrico Glaab, Anaïs Baudot, Natalio Krasnogor, and Alfonso Valencia. TopoGSA: network topological gene set analysis. *Bioinformatics (Oxford, England)*, 26(9):1271–1272, May 2010. PMID: 20335277.

[86] Mark Goadrich, Louis Oliphant, and Jude Shavlik. Learning ensembles of First-Order clauses for Recall-Precision curves: A case study in biomedical information extraction. In *Inductive Logic Programming*, pages 421–456, 2004.

[87] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690, May 2007. PMID: 17502601.

[88] A Green, A Svejgaard, P Platz, L P Ryder, B K Jakobsen, N E Morton, and C J MacLean. The genetic susceptibility to insulin-dependent diabetes mellitus: combined segregation and linkage analysis. *Genetic Epidemiology*, 2(1):1–15, 1985. PMID: 3863777.

[89] Lorna Gregory, Paul J Came, and Stephen Brown. Stem cell regulation by JAK/STAT signaling in drosophila. *Seminars in Cell & Developmental Biology*, 19(4):407–413, August 2008. PMID: 18603010.

[90] Kristin C Gunsalus, Hui Ge, Aaron J Schetter, Debra S Goldberg, Jing-Dong J Han, Tong Hao, Gabriel F Berriz, Nicolas Bertin, Jerry Huang, Ling-Shiang Chuang, Ning Li, Ramamurthy Mani, Anthony A Hyman, Birte Sönnichsen, Christophe J Echeverri, Frederick P Roth, Marc Vidal, and Fabio Piano. Predictive models of molecular machines involved in caenorhabditis elegans early embryogenesis. *Nature*, 436(7052):861–865, August 2005. PMID: 16094371.

[91] Syed Haider, Benoit Ballester, Damian Smedley, Junjun Zhang, Peter Rice, and Arek Kasprzyk. BioMart central portal–unified access to biological data. *Nucleic Acids Research*, 37(Web Server issue):W23–27, July 2009. PMID: 19420058.

[92] R W Haile, S E Hodge, B R Visscher, M A Spence, R Detels, T L McAuliffe, M S Park, and J P Dudley. Genetic susceptibility to multiple sclerosis: a linkage analysis with age-of-onset corrections. *Clinical Genetics*, 18(3):160–167, September 1980. PMID: 7438496.

[93] R W Haile, L Iselius, S E Hodge, N E Morton, and R Detels. Segregation and linkage analysis of 40 multiplex multiple sclerosis families. *Human Heredity*, 31(4):252–258, 1981. PMID: 7287016.

[94] Ada Hamosh, Alan F Scott, Joanna Amberger, Carol Bocchini, David Valle, and Victor A McKusick. Online mendelian inheritance in man (OMIM), a

knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, January 2002. PMID: 11752252.

[95] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–517, January 2005. PMID: 15608251.

[96] E L Harris, D K Wagener, J S Dorman, and A L Drash. Detection of genetic heterogeneity between families of insulin-dependent diabetes mellitus patients using linkage analysis. *American Journal of Human Genetics*, 37(1):102–113, January 1985. PMID: 3856383.

[97] Michael A Hauser, Yi-Ju Li, Satoshi Takeuchi, Robert Walters, Maher Noureddine, Melinda Maready, Tiffany Darden, Christine Hulette, Eden Martin, Elizabeth Hauser, Hong Xu, Don Schmechel, Judith E Stenger, Fred Dietrich, and Jeffery Vance. Genomic convergence: identifying candidate genes for parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Human Molecular Genetics*, 12(6):671–677, March 2003. PMID: 12620972.

[98] John Hawkins, Charles Grant, William Stafford Noble, and Timothy L Bailey. Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics (Oxford, England)*, 25(12):i339–347, June 2009. PMID: 19478008.

[99] Ruth Van Hellemont, Pieter Monsieurs, Gert Thijs, Bart de Moor, Yves Van de Peer, and Kathleen Marchal. A novel approach to identifying regulatory motifs in distantly related genomes. *Genome Biology*, 6(13):R113, 2005. PMID: 16420672.

[100] Julien I E Hoffman and Samuel Kaplan. The incidence of congenital heart disease. *Journal of the American College of Cardiology*, 39(12):1890–1900, June 2002. PMID: 12084585.

[101] Robert Hoffmann. A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9):1047–1051, September 2008. PMID: 18728691.

[102] Dimitar Hristovski, Borut Peterlin, Joyce A Mitchell, and Susanne M Humphrey. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2-4):289–298, March 2005. PMID: 15694635.

[103] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009. PMID: 19131956.

[104] Hailiang Huang and Joel S Bader. Precision and recall estimates for two-hybrid screens. *Bioinformatics (Oxford, England)*, 25(3):372–378, February 2009. PMID: 19091773.

[105] Hailiang Huang, Bruno M Jedynak, and Joel S Bader. Where have all the interactions gone? estimating the coverage of two-hybrid protein interaction maps. *PLoS Computational Biology*, 3(11):e214, November 2007. PMID: 18039026.

[106] Hui Huang, Eitan E Winter, Huajun Wang, Keith G Weinstock, Heming Xing, Leo Goodstadt, Peter D Stenson, David N Cooper, Douglas Smith, M Mar Albà, Chris P Ponting, and Kim Fechtel. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biology*, 5(7):R47, 2004. PMID: 15239832.

[107] Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web Server issue):W729–732, July 2006. PMID: 16845108.

[108] Curtis Huttenhower, K Tsheko Mutungu, Natasha Indik, Woongcheol Yang, Mark Schroeder, Joshua J Forman, Olga G Troyanskaya, and Hilary A Coller. Detailing regulatory networks through large scale data integration. *Bioinformatics (Oxford, England)*, 25(24):3267–3274, December 2009. PMID: 19825796.

[109] G Hutvágner, J McLachlan, A E Pasquinelli, E Bálint, T Tuschl, and P D Zamore. A cellular function for the RNA-interference enzyme dicer in the maturation of the let-7 small temporal RNA. *Science (New York, N.Y.)*, 293(5531):834–838, August 2001. PMID: 11452083.

[110] Janna E Hutz, Aldi T Kraja, Howard L McLeod, and Michael A Province. CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genetic Epidemiology*, 32(8):779–790, December 2008. PMID: 18613097.

[111] M Iizuka, Y Kubo, I Tsunenari, C X Pan, I Akiba, and T Kono. Functional characterization and localization of a cardiac-type inwardly rectifying k+ channel. *Receptors & Channels*, 3(4):299–315, 1995. PMID: 8834003.

[112] Md Shahidul Islam and Du Toit Loots. Experimental rodent models of type 2 diabetes: a review. *Methods and Findings in Experimental and Clinical Pharmacology*, 31(4):249–261, May 2009. PMID: 19557203.

[113] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein

interactions from genomic data. *Science (New York, N.Y.)*, 302(5644):449–453, October 2003. PMID: 14564010.

[114] Kathy J Jenkins, Adolfo Correa, Jeffrey A Feinstein, Lorenzo Botto, Amy E Britt, Stephen R Daniels, Marsha Elixson, Carole A Warnes, and Catherine L Webb. 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*, 115(23):2995–3014, June 2007. PMID: 17519397.

[115] G Jimenez-Sanchez, B Childs, and D Valle. Human disease genes. *Nature*, 409(6822):853–855, February 2001. PMID: 11237009.

[116] G Joshi-Tope, M Gillespie, I Vastrik, P D'Eustachio, E Schmidt, B de Bono, B Jassal, G R Gopinath, G R Wu, L Matthews, S Lewis, E Birney, and L Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue):D428–432, January 2005. PMID: 15608231.

[117] Jaehee Jung, Gangman Yi, Serenella A Sukno, and Michael R Thon. PoGO: prediction of gene ontology terms for fungal proteins. *BMC Bioinformatics*, 11:215, 2010. PMID: 20429880.

[118] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000. PMID: 10592173.

[119] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(Database issue):D355–360, January 2010. PMID: 19880382.

[120] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue):D354–357, January 2006. PMID: 16381885.

[121] Vicky Katsanou, Stavros Milatos, Anthie Yiakouvaki, Nikos Sgantzis, Anastasia Kotsoni, Maria Alexiou, Vaggelis Harokopos, Vassilis Aidinis, Myriam Hemberger, and Dimitris L Kontoyiannis. The RNA-binding protein Elavl1/HuR is essential for placental branching morphogenesis and embryonic development. *Molecular and Cellular Biology*, 29(10):2762–2776, May 2009. PMID: 19307312.

[122] R F Ketting, S E Fischer, E Bernstein, T Sijen, G J Hannon, and R H Plasterk. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in c. elegans. *Genes & Development*, 15(20):2654–2659, October 2001. PMID: 11641272.

[123] Sang-Bae Kim, Sungjin Yang, Seon-Kyu Kim, Sang Cheol Kim, Hyun Goo Woo, David J Volsky, Seon-Young Kim, and In-Sun Chu. GAzer: gene set analyzer. *Bioinformatics (Oxford, England)*, 23(13):1697–1699, July 2007. PMID: 17468122.

[124] Sang-Woon Kim and B John Oommen. On using prototype reduction schemes and classifier fusion strategies to optimize kernel-based nonlinear subspace methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):455–460, March 2005. PMID: 15747799.

[125] K Kleppe, E Ohtsuka, R Kleppe, I Molineux, and H G Khorana. Studies on polynucleotides. XCVI. repair replications of short synthetic DNA's as catalyzed by DNA polymerases. *Journal of Molecular Biology*, 56(2):341–361, March 1971. PMID: 4927950.

[126] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82(4):949–958, April 2008. PMID: 18371930.

[127] Stanley Kok and Pedro Domingos. Learning the structure of markov logic networks. In *Proceedings of the 22nd international conference on Machine learning*, pages 441–448, Bonn, Germany, 2005. ACM.

[128] Kasper Lage, E Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, Yves Moreau, and Søren Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316, March 2007. PMID: 17344885.

[129] M Lagos-Quintana, R Rauhut, W Lendeckel, and T Tuschl. Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294(5543):853–858, October 2001. PMID: 11679670.

[130] Gert R G Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics (Oxford, England)*, 20(16):2626–2635, November 2004. PMID: 15130933.

[131] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French,

D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, and J Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. PMID: 11237011.

[132] D A Lashkari, J L DeRisi, J H McCusker, A F Namath, C Gentile, S Y Hwang, P O Brown, and R W Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13057–13062, November 1997. PMID: 9371799.

[133] Natalie Lavine, Nathalie Ethier, James N Oak, Lin Pei, Fang Liu, Phan Trieu, R Victor Rebois, Michel Bouvier, Terence E Hebert, and Hubert H M Van Tol. G protein-coupled receptors form stable complexes with inwardly

rectifying potassium channels and adenylyl cyclase. *The Journal of Biological Chemistry*, 277(48):46010–46019, November 2002. PMID: 12297500.

[134] Sonia Leach, Aaron Gabow, Lawrence Hunter, and Debra S Goldberg. Assessing and combining reliability of protein interaction sources. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 433–444, 2007. PMID: 17990508.

[135] Phillip Lee, Eileen P Treacy, Eric Crombez, Melissa Wasserstein, Lewis Waber, Jon Wolff, Udo Wendel, Alex Dorenbaum, Judith Bebchuk, Heidi Christ-Schmidt, Margretta Seashore, Marcello Giovannini, Barbara K Burton, and Andrew A Morris. Safety and efficacy of 22 weeks of treatment with sapropterin dihydrochloride in patients with phenylketonuria. *American Journal of Medical Genetics. Part A*, 146A(22):2851–2859, November 2008. PMID: 18932221.

[136] Karen Lemmens, Tijl De Bie, Thomas Dhollander, Sigrid C De Keersmaecker, Inge M Thijs, Geert Schoofs, Ami De Weerdt, Bart De Moor, Jos Vanderleyden, Julio Collado-Vides, Kristof Engelen, and Kathleen Marchal. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in escherichia coli. *Genome Biology*, 10(3):R27, 2009. PMID: 19265557.

[137] Harvey L Levy, Andrzej Milanowski, Anupam Chakrapani, Maureen Cleary, Philip Lee, Friedrich K Trefz, Chester B Whitley, François Feillet, Annette S Feigenbaum, Judith D Bebchuk, Heidi Christ-Schmidt, and Alex Dorenbaum. Efficacy of sapropterin dihydrochloride (tetrahydrobiopterin, 6R-BH4) for reduction of phenylalanine concentration in patients with phenylketonuria: a phase III randomised placebo-controlled study. *Lancet*, 370(9586):504–510, August 2007. PMID: 17693179.

[138] Li Li, Jianzhen Xu, Deyin Yang, Xiaorong Tan, and Hongfei Wang. Computational approaches for microRNA studies: a review. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 21(1-2):1–12, February 2010. PMID: 20012966.

[139] Xiao-Gang Liu, Yong-Jun Liu, Jianfeng Liu, Yufang Pei, Dong-Hai Xiong, Hui Shen, Hong-Yi Deng, Christopher J Papasian, Betty M Drees, James J Hamilton, Robert R Recker, and Hong-Wen Deng. A bivariate whole genome linkage study identified genomic regions influencing both BMD and bone structure. *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research*, 23(11):1806–1814, November 2008. PMID: 18597637.

[140] Xiaowen Liu and Lusheng Wang. Computing the maximum similarity bi-clusters of gene expression data. *Bioinformatics (Oxford, England)*, 23(1):50–56, January 2007. PMID: 17090578.

[141] Peter Van Loo and Peter Marynen. Computational methods for the detection of cis-regulatory modules. *Briefings in Bioinformatics*, 10(5):509–524, September 2009. PMID: 19498042.

[142] Núria López-Bigas and Christos A Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Research*, 32(10):3108–3114, 2004. PMID: 15181176.

[143] Chenqi Lu, Xiaohua Hu, Guiying Wang, L J Leach, Shengjie Yang, M J Kearsey, and Z W Luo. Why do essential proteins tend to be clustered in the yeast interactome network? *Molecular bioSystems*, 6(5):871–877, May 2010. PMID: 20567773.

[144] Long J Lu, Yu Xia, Alberto Paccanaro, Haiyuan Yu, and Mark Gerstein. Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15(7):945–953, July 2005. PMID: 15998909.

[145] Feng Luo, Yunfeng Yang, Jianxin Zhong, Haichun Gao, Latifur Khan, Dorothea K Thompson, and Jizhong Zhou. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics*, 8:299, 2007. PMID: 17697349.

[146] James R Lupski, Jeffrey G Reid, Claudia Gonzaga-Jauregui, David Rio Deiros, David C Y Chen, Lynne Nazareth, Matthew Bainbridge, Huyen Dinh, Chyn Jing, David A Wheeler, Amy L McGuire, Feng Zhang, Pawel Stankiewicz, John J Halperin, Chengyong Yang, Curtis Gehman, Danwei Guo, Rola K Irikat, Warren Tom, Nick J Fantin, Donna M Muzny, and Richard A Gibbs. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England Journal of Medicine*, 362(13):1181–1191, April 2010. PMID: 20220177.

[147] Jan Luts, Fabian Ojeda, Raf Van de Plas, Bart De Moor, Sabine Van Huffel, and Johan A K Suykens. A tutorial on support vector machine-based methods for classification problems in chemometrics. *Analytica Chimica Acta*, 665(2):129–145, April 2010. PMID: 20417323.

[148] Xiaotu Ma, Hyunju Lee, Li Wang, and Fengzhu Sun. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics (Oxford, England)*, 23(2):215–221, January 2007. PMID: 17098772.

[149] Nicole K MacLennan, Jun Dong, Jason E Aten, Steve Horvath, Lola Rahib, Loren Ornelas, Katrina M Dipple, and Edward R B McCabe. Weighted gene co-expression network analysis identifies biomarkers in glycerol kinase deficient mice. *Molecular Genetics and Metabolism*, 98(1-2):203–214, October 2009. PMID: 19546021.

[150] Ani Manichaikul, Lila Ghamsari, Erik F Y Hom, Chenwei Lin, Ryan R Murray, Roger L Chang, S Balaji, Tong Hao, Yun Shen, Arvind K Chavali, Ines Thiele, Xinping Yang, Changyu Fan, Elizabeth Mello, David E Hill, Marc Vidal, Kourosh Salehi-Ashtiani, and Jason A Papin. Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nature Methods*, 6(8):589–592, August 2009. PMID: 19597503.

[151] Nicky Manning and Nick Archer. A study to determine the incidence of structural congenital heart disease in monochorionic twins. *Prenatal Diagnosis*, 26(11):1062–1064, November 2006. PMID: 16958142.

[152] Daniele Masotti, Christine Nardini, Simona Rossi, Elena Bonora, Giovanni Romeo, Stefano Volinia, and Luca Benini. TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders. *Bioinformatics (Oxford, England)*, 24(3):428–429, February 2008. PMID: 18048394.

[153] Suresh Mathivanan, Balamurugan Periaswamy, T K B Gandhi, Kumaran Kandasamy, Shubha Suresh, Riaz Mohmood, Y L Ramachandra, and Akhilesh Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7 Suppl 5:S19, 2006. PMID: 17254303.

[154] Cédric Maurange and Renato Paro. A cellular memory module conveys epigenetic inheritance of hedgehog expression during drosophila wing imaginal disc development. *Genes & Development*, 16(20):2672–2683, October 2002. PMID: 12381666.

[155] V A McKUSICK. The cardiovascular aspects of marfan's syndrome: a heritable disorder of connective tissue. *Circulation*, 11(3):321–342, March 1955. PMID: 14352380.

[156] Victor A McKusick. *Mendelian inheritance in man: catalogs of autosomal dominant, autosomal recessive and X-linked phenotypes.* Johns Hopkins University Press, Baltimore, 1966.

[157] Victor A McKusick. Mendelian inheritance in man and its online version, OMIM. *American Journal of Human Genetics*, 80(4):588–604, April 2007. PMID: 17357067.

[158] Suyu Mei and Wang Fei. Amino acid classification based spectrum kernel fusion for protein subnuclear localization. *BMC Bioinformatics*, 11 Suppl 1:S17, 2010. PMID: 20122188.

[159] M Milán, U Weihe, S Tiong, W Bender, and S M Cohen. msh specifies dorsal cell fate in the drosophila wing. *Development (Cambridge, England)*, 128(17):3263–3268, September 2001. PMID: 11546743.

[160] Gopa R Mishra, M Suresh, K Kumaran, N Kannabiran, Shubha Suresh, P Bala, K Shivakumar, N Anuradha, Raghunath Reddy, T Madhan Raghavan, Shalini Menon, G Hanumanthu, Malvika Gupta, Sapna Upendran, Shweta Gupta, M Mahesh, Bincy Jacob, Pinky Mathew, Pritam Chatterjee, K S Arun, Salil Sharma, K N Chandrika, Nandan Deshpande, Kshitish Palvankar, R Raghavnath, R Krishnakanth, Hiren Karathia, B Rekha, Rashmi Nayak, G Vishnupriya, H G Mohan Kumar, M Nagini, G S Sameer Kumar, Rojan Jose, P Deepthi, S Sujatha Mohan, T K B Gandhi, H C Harsha, Krishna S Deshpande, Malabika Sarker, T S Keshava Prasad, and Akhilesh Pandey. Human protein reference database–2006 update. *Nucleic Acids Research*, 34(Database issue):D411–414, January 2006. PMID: 16381900.

[161] Pieter Monsieurs, Gert Thijs, Abeer A Fadda, Sigrid C J De Keersmaecker, Jozef Vanderleyden, Bart De Moor, and Kathleen Marchal. More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics*, 7:160, 2006. PMID: 16549017.

[162] Fantine Mordelet and Jean-Philippe Vert. SIRENE: supervised inference of regulatory networks. *Bioinformatics (Oxford, England)*, 24(16):i76–82, August 2008. PMID: 18689844.

[163] Julie L Morrison, Rainer Breitling, Desmond J Higham, and David R Gilbert. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, 6:233, 2005. PMID: 16176585.

[164] J Moult, T Hubbard, S H Bryant, K Fidelis, and J T Pedersen. Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins*, Suppl 1:2–6, 1997. PMID: 9485489.

[165] Christopher J Mungall, Colin Batchelor, and Karen Eilbeck. Evolution of the sequence ontology terms and relationships. *Journal of Biomedical Informatics*, March 2010. PMID: 20226267.

[166] Carl Murie, Owen Woody, Anna Y Lee, and Robert Nadon. Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics*, 10:45, 2009. PMID: 19192265.

[167] Chad L Myers, Daniel R Barrett, Matthew A Hibbs, Curtis Huttenhower, and Olga G Troyanskaya. Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7:187, 2006. PMID: 16869964.

[168] A C Nicholls, J E Oliver, S McCarron, J B Harrison, D S Greenspan, and F M Pope. An exon skipping mutation of a type v collagen gene (COL5A1) in Ehlers-Danlos syndrome. *Journal of Medical Genetics*, 33(11):940–946, November 1996. PMID: 8950675.

[169] Daniela Nitsch, Léon-Charles Tranchevent, Bernard Thienpont, Lieven Thorrez, Hilde Van Esch, Koenraad Devriendt, and Yves Moreau. Network analysis of differential expression for the identification of disease-causing genes. *PloS One*, 4(5):e5526, 2009. PMID: 19436755.

[170] Kevin P O'Brien, Maido Remm, and Erik L L Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33(Database issue):D476–480, January 2005. PMID: 15608241.

[171] E O'Hara, B Cohen, S M Cohen, and W McGinnis. Distal-less is a downstream gene of deformed required for ventral maxillary identity. *Development (Cambridge, England)*, 117(3):847–856, March 1993. PMID: 8100764.

[172] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat, and Peter Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics (Oxford, England)*, 20(17):3045–3054, November 2004. PMID: 15201187.

[173] Fabian Ojeda, Johan A K Suykens, and Bart De Moor. Low rank updated LS-SVM classifiers for fast variable selection. *Neural Networks: The Official Journal of the International Neural Network Society*, 21(2-3):437–449, April 2008. PMID: 18343309.

[174] Gabriel Ostlund, Thomas Schmitt, Kristoffer Forslund, Tina Köstler, David N Messina, Sanjit Roopra, Oliver Frings, and Erik L L Sonnhammer. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(Database issue):D196–203, January 2010. PMID: 19892828.

[175] Martin Oti, Jeroen van Reeuwijk, Martijn A Huynen, and Han G Brunner. Conserved co-expression for candidate disease gene prioritization. *BMC Bioinformatics*, 9:208, 2008. PMID: 18433471.

[176] PA Owren. The coagulation of blood: investigations of a new clotting factor. *Acta Med Scand*, 128[Suppl](1), 1947.

[177] Subhamoy Pal and Louisa P Wu. Pattern recognition receptors in the fly: lessons we can learn from the drosophila melanogaster immune system. *Fly*, 3(2):121–129, June 2009. PMID: 19440043.

[178] Wei Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics (Oxford, England)*, 18(4):546–554, April 2002. PMID: 12016052.

[179] Inju Park, Seong-Eui Hong, Tae-Wan Kim, Jiae Lee, Jungsu Oh, Eunyoung Choi, Cecil Han, Hoyong Lee, Do Han Kim, and Chunghee Cho.

Comprehensive identification and characterization of novel cardiac genes in mouse. *Journal of Molecular and Cellular Cardiology*, 43(2):93–106, August 2007. PMID: 17599348.

[180] Jeehye Park, Yongsung Kim, and Jongkyeong Chung. Mitochondrial dysfunction and parkinson's disease genes: insights from drosophila. *Disease Models & Mechanisms*, 2(7-8):336–340, August 2009. PMID: 19553694.

[181] Helen Parkinson, Misha Kapushesky, Nikolay Kolesnikov, Gabriella Rustici, Mohammad Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Ele Holloway, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Tim F Rayner, Faisal Rezwan, Anjan Sharma, Eleanor Williams, Xiangqun Zheng Bradley, Tomasz Adamusiak, Marco Brandizi, Tony Burdett, Richard Coulson, Maria Krestyaninova, Pavel Kurnosov, Eamonn Maguire, Sudeshna Guha Neogi, Philippe Rocca-Serra, Susanna-Assunta Sansone, Nataliya Sklyar, Mengyao Zhao, Ugis Sarkans, and Alvis Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(Database issue):D868–872, January 2009. PMID: 19015125.

[182] A E Pasquinelli, B J Reinhart, F Slack, M Q Martindale, M I Kuroda, B Maller, D C Hayward, E E Ball, B Degnan, P Müller, J Spring, A Srinivasan, M Fishman, J Finnerty, J Corbo, M Levine, P Leahy, E Davidson, and G Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, November 2000. PMID: 11081512.

[183] Topon Kumar Paul and Hitoshi Iba. Gene selection for classification of cancers using probabilistic model building genetic algorithm. *Bio Systems*, 82(3):208–225, December 2005. PMID: 16112804.

[184] W R Pearson and D J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448, April 1988. PMID: 3162770.

[185] Carolina Perez-Iratxeta, Peer Bork, and Miguel A Andrade. Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31(3):316–319, July 2002. PMID: 12006977.

[186] Carolina Perez-Iratxeta, Peer Bork, and Miguel A Andrade-Navarro. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Research*, 35(Web Server issue):W212–216, July 2007. PMID: 17478516.

[187] Carolina Perez-Iratxeta, Matthias Wjst, Peer Bork, and Miguel A Andrade. G2D: a tool for mining genes associated with disease. *BMC Genetics*, 6:45, 2005. PMID: 16115313.

[188] Suraj Peri, J Daniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, T K B Gandhi, Mads Gronborg, Nieves Ibarrola, Nandan Deshpande, K Shanker, H N Shivashankar, B P Rashmi, M A Ramya, Zhixing Zhao, K N Chandrika, N Padma, H C Harsha, A J Yatish, M P Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Shubha Suresh, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Subhalakshmi Krishna, Mary Joy, Sanjeev K Anand, V Madavan, Ansamma Joseph, Guang W Wong, William P Schiemann, Stefan N Constantinescu, Lily Huang, Roya Khosravi-Far, Hanno Steen, Muneesh Tewari, Saghi Ghaffari, Gerard C Blobe, Chi V Dang, Joe G N Garcia, Jonathan Pevsner, Ole N Jensen, Peter Roepstorff, Krishna S Deshpande, Arul M Chinnaiyan, Ada Hamosh, Aravinda Chakravarti, and Akhilesh Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, October 2003. PMID: 14525934.

[189] George H Perry, Nathaniel J Dominy, Katrina G Claw, Arthur S Lee, Heike Fiegler, Richard Redon, John Werner, Fernando A Villanea, Joanna L Mountain, Rajeev Misra, Nigel P Carter, Charles Lee, and Anne C Stone. Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39(10):1256–1260, October 2007. PMID: 17828263.

[190] Mary Ella Pierpont, Craig T Basson, D Woodrow Benson, Bruce D Gelb, Therese M Giglia, Elizabeth Goldmuntz, Glenn McGee, Craig A Sable, Deepak Srivastava, and Catherine L Webb. 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*, 115(23):3015–3038, June 2007. PMID: 17519398.

[191] Martin Poot, Marc J Eleveld, Ruben van 't Slot, Hans Kristian Ploos van Amstel, and Ron Hochstenbach. Recurrent copy number changes in mentally retarded children harbour genes involved in cellular localization and the glutamate receptor complex. *European Journal of Human Genetics: EJHG*, 18(1):39–46, January 2010. PMID: 19623214.

[192] Maximilian G Posch, Andreas Perrot, Katharina Schmitt, Sebastian Mittelhaus, Eva-Maria Esenwein, Brigitte Stiller, Christian Geier, Rainer Dietz, Reinhard Gessner, Cemil Ozcelik, and Felix Berger. Mutations in GATA4, NKX2.5, CRELD1, and BMP4 are infrequently found in patients with congenital cardiac septal defects. *American Journal of Medical Genetics. Part A*, 146A(2):251–253, January 2008. PMID: 18076106.

[193] T S Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla,

Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C J Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y L Ramachandra, V Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–772, January 2009. PMID: 18988627.

[194] Rainer Pudimat, Ernst-Günter Schukat-Talamazzini, and Rolf Backofen. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics (Oxford, England)*, 21(14):3082–3088, July 2005. PMID: 15905283.

[195] Y Qiao, N Riendeau, M Koochek, X Liu, Chansonette Harvard, M J Hildebrand, J J A Holden, E Rajcan-Separovic, and M E S Lewis. Phenomic determinants of genomic variation in autism spectrum disorders. *Journal of Medical Genetics*, 46(10):680–688, October 2009. PMID: 19625284.

[196] Predrag Radivojac, Kang Peng, Wyatt T Clark, Brandon J Peters, Amrita Mohan, Sean M Boyle, and Sean D Mooney. An integrated approach to inferring gene-disease associations in humans. *Proteins*, 72(3):1030–1037, August 2008. PMID: 18300252.

[197] Fidel Ramírez, Andreas Schlicker, Yassen Assenov, Thomas Lengauer, and Mario Albrecht. Computational analysis of human protein interaction networks. *Proteomics*, 7(15):2541–2552, August 2007. PMID: 17647236.

[198] Thomas Rattei, Roland Arnold, Patrick Tischler, Dominik Lindner, Volker Stümpflen, and H Werner Mewes. SIMAP: the similarity matrix of proteins. *Nucleic Acids Research*, 34(Database issue):D252–256, January 2006. PMID: 16381858.

[199] Thomas Rattei, Patrick Tischler, Roland Arnold, Franz Hamberger, Jörg Krebs, Jan Krumsiek, Benedikt Wachinger, Volker Stümpflen, and Werner Mewes. SIMAP–structuring the network of protein similarities. *Nucleic Acids Research*, 36(Database issue):D289–292, January 2008. PMID: 18037617.

[200] Richard Redon, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, Eun Kyung Cho, Stephanie Dallaire, Jennifer L Freeman, Juan R González, Mònica Gratacòs, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R MacDonald, Christian R Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol,

Donald F Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P Carter, Hiroyuki Aburatani, Charles Lee, Keith W Jones, Stephen W Scherer, and Matthew E Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, November 2006. PMID: 17122850.

[201] B J Reinhart, F J Slack, M Basson, A E Pasquinelli, J C Bettinger, A E Rougvie, H R Horvitz, and G Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–906, February 2000. PMID: 10706289.

[202] Catherine Bowes Rickman, Jessica N Ebright, Zachary J Zavodni, Ling Yu, Tianyuan Wang, Stephen P Daiger, Graeme Wistow, Kathy Boon, and Michael A Hauser. Defining the human macula transcriptome and candidate retinal disease genes using EyeSAGE. *Investigative Ophthalmology & Visual Science*, 47(6):2305–2316, June 2006. PMID: 16723438.

[203] J R Riordan, J M Rommens, B Kerem, N Alon, R Rozmahel, Z Grzelczak, J Zielenski, S Lok, N Plavsic, and J L Chou. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science (New York, N.Y.)*, 245(4922):1066–1073, September 1989. PMID: 2475911.

[204] Cornelius Rosse and José L V Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6):478–500, December 2003. PMID: 14759820.

[205] Simona Rossi, Daniele Masotti, Christine Nardini, Elena Bonora, Giovanni Romeo, Enrico Macii, Luca Benini, and Stefano Volinia. TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Research*, 34(Web Server issue):W285–292, July 2006. PMID: 16845011.

[206] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):D449–451, January 2004. PMID: 14681454.

[207] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael Dicuccio, Scott Federhen, Michael Feolo, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J Lipman, Zhiyong Lu, Thomas L Madden, Tom Madej, Donna R Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrachi, James Ostell, Anna Panchenko, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Stephen T Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A Tatusova, Lukas Wagner, Yanli Wang, W John Wilbur, Eugene Yaschenko, and Jian Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 38(Database issue):D5–16, January 2010. PMID: 19910364.

[208] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj Guhathakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, Pek Yee Lum, Amy Leonardson, Rolf Thieringer, Joseph M Metzger, Liming Yang, John Castle, Haoyuan Zhu, Shera F Kash, Thomas A Drake, Alan Sachs, and Aldons J Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, July 2005. PMID: 15965475.

[209] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–470, October 1995. PMID: 7569999.

[210] B Schölkopf, J C Platt, J Shawe-Taylor, A J Smola, and R C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, July 2001. PMID: 11440593.

[211] Dominik Seelow, Jana Marie Schwarz, and Markus Schuelke. GeneDistiller–distilling candidate genes from linkage intervals. *PloS One*, 3(12):e3874, 2008. PMID: 19057649.

[212] Urmi Sengupta, Sanchaita Ukil, Nevenka Dimitrova, and Shipra Agrawal. Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications. *PloS One*, 4(12):e8100, 2009. PMID: 19997558.

[213] Jerzy Silberring and Pawel Ciborowski. Biomarker discovery and clinical proteomics. *Trends in Analytical Chemistry: TRAC*, 29(2):128, February 2010. PMID: 20174458.

[214] Diego G Silva, Christian Schönbach, Vladimir Brusic, Luis A Socha, Takeshi Nagashima, and Nikolai Petrovsky. Identification of "pathologs" (disease-related genes) from the RIKEN mouse cDNA dataset using human curation plus FACTS, a new biological information extraction system. *BMC Genomics*, 5(1):28, April 2004. PMID: 15115540.

[215] Parag Singla and Pedro Domingos. Discriminative training of markov logic networks. *Proceedings of the National Conference on Artificial Intelligence*, 2005.

[216] Damian Smedley, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. BioMart–biological queries made easy. *BMC Genomics*, 10:22, 2009. PMID: 19144180.

[217] Cynthia L Smith and Janan T Eppig. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 1(3):390–399, 2009. PMID: 20052305.

[218] Cynthia L Smith, Carroll-Ann W Goldsmith, and Janan T Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1):R7, 2005. PMID: 15642099.

[219] Nick G C Smith and Adam Eyre-Walker. Human disease genes: patterns and predictions. *Gene*, 318:169–175, October 2003. PMID: 14585509.

[220] Hon-Cheong So, Pui Y Fong, Ronald Y L Chen, Tomy C K Hui, Mandy Y M Ng, Stacey S Cherny, William W M Mak, Eric F C Cheung, Raymond C K Chan, Eric Y H Chen, Tao Li, and Pak C Sham. Identification of neuroglycan c and interacting partners as potential susceptibility genes for schizophrenia in a southern chinese population. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, 153B(1):103–113, January 2010. PMID: 19367581.

[221] Silvia Sookoian, Tomas Fernandez Gianotti, Carolina Gemma, Adriana L Burgueño, and Carlos J Pirola. Role of genetic variation in insulin-like growth factor 1 receptor on insulin resistance and arterial hypertension. *Journal of Hypertension*, 28(6):1194–1202, June 2010. PMID: 20179633.

[222] Silvia Sookoian, Tomas Fernández Gianotti, Mariano Schuman, and Carlos Jose Pirola. Gene prioritization based on biological plausibility over genome wide association studies renders new loci associated with type 2 diabetes. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 11(5):338–343, May 2009. PMID: 19346957.

[223] E M Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98(3):503–517, November 1975. PMID: 1195397.

[224] E Storey, M Bahlo, M Fahey, O Sisson, C J Lueck, and R J M Gardner. A new dominantly inherited pure cerebellar ataxia, SCA 30. *Journal of Neurology, Neurosurgery, and Psychiatry*, 80(4):408–411, April 2009. PMID: 18996908.

[225] Bing-Yu Sun, Xiao-Ming Zhang, Jiuyong Li, and Xue-Min Mao. Feature fusion using locally linear embedding for classification. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*, 21(1):163–168, January 2010. PMID: 19963695.

[226] Aik Choon Tan and David Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2(3 Suppl):S75–83, 2003. PMID: 15130820.

[227] Daisuke Tanaka, Kazuaki Nagashima, Mayumi Sasaki, Chizumi Yamada, Shogo Funakoshi, Kimiyo Akitomo, Katsunobu Takenaka, Kouji Harada, Akio Koizumi, and Nobuya Inagaki. GCKR mutations in japanese families with clustered type 2 diabetes. *Molecular Genetics and Metabolism*, 102(4):453–460, April 2011. PMID: 21236713.

[228] James Taylor, Ian Schenck, Dan Blankenberg, and Anton Nekrutenko. Using galaxy to perform large-scale interactive data analyses. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al*, Chapter 10:Unit 10.5, September 2007. PMID: 18428782.

[229] Erdahl T Teber, Jason Y Liu, Sara Ballouz, Diane Fatkin, and Merridee A Wouters. 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, 2009. PMID: 19208173.

[230] Christian Thiel, Kristin Kessler, Andreas Giessl, Arno Dimmler, Stavit A. Shalev, Sigrun von der Haar, Martin Zenker, Diana Zahnleiter, Hartmut Stöss, Ernst Beinder, Rami Abou Jamra, Arif B. Ekici, Nadja Schröder-Kreß, Thomas Aigner, Thomas Kirchner, André Reis, Johann H. Brandstätter, and Anita Rauch. NEK1 mutations cause Short-Rib polydactyly syndrome type majewski. *The American Journal of Human Genetics*, 88(1):106–114, January 2011.

[231] Bernard Thienpont, Luc Mertens, Thomy de Ravel, Benedicte Eyskens, Derize Boshoff, Nicole Maas, Jean-Pierre Fryns, Marc Gewillig, Joris R Vermeesch, and Koen Devriendt. Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. *European Heart Journal*, 28(22):2778–2784, November 2007. PMID: 17384091.

[232] Bernard Thienpont, Litu Zhang, Alex V Postma, Jeroen Breckpot, Léon-Charles Tranchevent, Peter Van Loo, Kjeld Møllgård, Niels Tommerup, Iben Bache, Zeynep Tümer, Klaartje van Engelen, Björn Menten, Geert Mortier, Darrel Waggoner, Marc Gewillig, Yves Moreau, Koen Devriendt, and Lars Allan Larsen. Haploinsufficiency of TAB2 causes congenital heart defects in humans. *American Journal of Human Genetics*, May 2010. PMID: 20493459.

[233] Thomas Thom, Nancy Haase, Wayne Rosamond, Virginia J Howard, John Rumsfeld, Teri Manolio, Zhi-Jie Zheng, Katherine Flegal, Christopher O'Donnell, Steven Kittner, Donald Lloyd-Jones, David C Goff, Yuling Hong, Robert Adams, Gary Friday, Karen Furie, Philip Gorelick, Brett Kissela, John Marler, James Meigs, Veronique Roger, Stephen Sidney, Paul Sorlie, Julia Steinberger, Sylvia Wasserthiel-Smoller, Matthew Wilson, and Philip Wolf. Heart disease and stroke statistics–2006 update: a report from the american

heart association statistics committee and stroke statistics subcommittee. *Circulation*, 113(6):e85–151, February 2006. PMID: 16407573.

[234] Nicki Tiffin, Euan Adie, Frances Turner, Han G Brunner, Marc A van Driel, Martin Oti, Nuria Lopez-Bigas, Christos Ouzounis, Carolina Perez-Iratxeta, Miguel A Andrade-Navarro, Adebowale Adeyemo, Mary Elizabeth Patti, Colin A M Semple, and Winston Hide. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Research*, 34(10):3067–3081, 2006. PMID: 16757574.

[235] Nicki Tiffin, Ikechi Okpechi, Carolina Perez-Iratxeta, Miguel A Andrade-Navarro, and Rajkumar Ramesar. Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes. *Physiological Genomics*, 35(1):55–64, September 2008. PMID: 18612082.

[236] Léon-Charles Tranchevent, Roland Barriot, Shi Yu, Steven Van Vooren, Peter Van Loo, Bert Coessens, Bart De Moor, Stein Aerts, and Yves Moreau. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Research*, 36(Web Server issue):W377–384, July 2008. PMID: 18508807.

[237] Abhishek Tripathi, Arto Klami, and Samuel Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008. PMID: 18291027.

[238] Olga G Troyanskaya, Kara Dolinski, Art B Owen, Russ B Altman, and David Botstein. A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8348–8353, July 2003. PMID: 12826619.

[239] Frances S Turner, Daniel R Clutterbuck, and Colin A M Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biology*, 4(11):R75, 2003. PMID: 14611661.

[240] Marc A van Driel, Jorn Bruggeman, Gert Vriend, Han G Brunner, and Jack A M Leunissen. A text-mining analysis of the human phenome. *European Journal of Human Genetics: EJHG*, 14(5):535–542, May 2006. PMID: 16493445.

[241] Geert Vandeweyer, Edwin Reyniers, Wim Wuyts, Liesbeth Rooms, and R Frank Kooy. CNV-WebStore: online CNV analysis, storage and interpretation. *BMC Bioinformatics*, 12:4, 2011. PMID: 21208430.

[242] Saran Vardhanabhuti, Steven J Blakemore, Steven M Clark, Sujoy Ghosh, Richard J Stephens, and Dilip Rajagopalan. A comparison of statistical

tests for detecting differential expression using affymetrix oligonucleotide microarrays. *Omics: A Journal of Integrative Biology*, 10(4):555–566, 2006. PMID: 17233564.

[243] T V Venkatesh and Harry B Harlow. Integromics: challenges in data integration. *Genome Biology*, 3(8):REPORTS4027, July 2002. PMID: 12186644.

[244] Raf Vervoort, Helga Ceulemans, Leen Van Aerschot, Rudi D'Hooge, and Guido David. Genetic modification of the inner ear lateral semicircular canal phenotype of the bmp4 haplo-insufficient mouse. *Biochemical and Biophysical Research Communications*, 394(3):780–785, April 2010. PMID: 20233579.

[245] Axel Visel, Christina Thaller, and Gregor Eichele. GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Research*, 32(Database issue):D552–556, January 2004. PMID: 14681479.

[246] Christian von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue):D433–437, January 2005. PMID: 15608232.

[247] E von Tschermack. Über künstliche kreuzung bei pisum sativum. *Berichte der Deutsche Botanische Gesellschaft*, 18:232–239, 1900.

[248] Steven Van Vooren, Bernard Thienpont, Björn Menten, Frank Speleman, Bart De Moor, Joris Vermeesch, and Yves Moreau. Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Research*, 35(8):2533–2543, 2007. PMID: 17403693.

[249] Albertha J M Walhout, Jérôme Reboul, Olena Shtanko, Nicolas Bertin, Philippe Vaglio, Hui Ge, Hongmei Lee, Lynn Doucette-Stamm, Kristin C Gunsalus, Aaron J Schetter, Diane G Morton, Kenneth J Kemphues, Valerie Reinke, Stuart K Kim, Fabio Piano, and Marc Vidal. Integrating interactome, phenome, and transcriptome mapping data for the c. elegans germline. *Current Biology: CB*, 12(22):1952–1958, November 2002. PMID: 12445390.

[250] S D Walter. The partial area under the summary ROC curve. *Statistics in Medicine*, 24(13):2025–2040, July 2005. PMID: 15900606.

[251] Kai Wang, Manikandan Narayanan, Hua Zhong, Martin Tompa, Eric E Schadt, and Jun Zhu. Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. *PLoS Computational Biology*, 5(12):e1000616, December 2009. PMID: 20019805.

[252] Taro Wasada. Adenosine triphosphate-sensitive potassium (K(ATP)) channel activity is coupled with insulin resistance in obesity and type 2 diabetes mellitus. *Internal Medicine (Tokyo, Japan)*, 41(2):84–90, February 2002. PMID: 11868613.

[253] John N Weinstein. Integromic analysis of the NCI-60 cancer cell lines. *Breast Disease*, 19:11–22, 2004. PMID: 15687693.

[254] R J Wenstrup, G T Langland, M C Willing, V N D'Souza, and W G Cole. A splice-junction mutation in the region of COL5A1 that codes for the carboxyl propeptide of pro alpha 1(V) chains results in the gravis form of the Ehlers-Danlos syndrome (type i). *Human Molecular Genetics*, 5(11):1733–1736, November 1996. PMID: 8923000.

[255] D L Wheeler, D M Church, A E Lash, D D Leipe, T L Madden, J U Pontius, G D Schuler, L M Schriml, T A Tatusova, L Wagner, and B A Rapp. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 29(1):11–16, January 2001. PMID: 11125038.

[256] M B White, J Amos, J M Hsu, B Gerrard, P Finn, and M Dean. A frame-shift mutation in the cystic fibrosis gene. *Nature*, 344(6267):665–667, April 1990. PMID: 1691449.

[257] R Whittemore, J A Wells, and X Castellsague. A second-generation study of 427 probands with congenital heart defects and their 837 children. *Journal of the American College of Cardiology*, 23(6):1459–1467, May 1994. PMID: 8176107.

[258] An Windelinckx, Robert Vlietinck, Jeroen Aerssens, Gaston Beunen, and Martine A I Thomis. Selection of genes and single nucleotide polymorphisms for fine mapping starting from a broad linkage region. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies*, 10(6):871–885, December 2007. PMID: 18179400.

[259] Graeme Wistow, Katherine Peterson, James Gao, Patee Buchoff, Cynthia Jaworski, Catherine Bowes-Rickman, Jessica N Ebright, Michael A Hauser, and David Hoover. NEIBank: genomics and bioinformatics resources for vision research. *Molecular Vision*, 14:1327–1337, 2008. PMID: 18648525.

[260] Sharyl L Wong, Lan V Zhang, Amy H Y Tong, Zhijian Li, Debra S Goldberg, Oliver D King, Guillaume Lesage, Marc Vidal, Brenda Andrews, Howard Bussey, Charles Boone, and Frederick P Roth. Combining biological networks to predict genetic interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(44):15682–15687, November 2004. PMID: 15496468.

[261] Randy Z Wu, Christina Chaivorapol, Jiashun Zheng, Hao Li, and Shoudan Liang. fREDUCE: detection of degenerate regulatory elements using correlation with expression. *BMC Bioinformatics*, 8:399, 2007. PMID: 17941998.

[262] I Xenarios, E Fernandez, L Salwinski, X J Duan, M J Thompson, E M Marcotte, and D Eisenberg. DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Research*, 29(1):239–241, January 2001. PMID: 11125102.

[263] I Xenarios, D W Rice, L Salwinski, M K Baron, E M Marcotte, and D Eisenberg. DIP: the database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291, January 2000. PMID: 10592249.

[264] Ioannis Xenarios, Lukasz Salwínski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305, January 2002. PMID: 11752321.

[265] Qing Xiong, Yuhui Qiu, and Weikuan Gu. PGMapper: a web-based tool linking phenotype to genes. *Bioinformatics (Oxford, England)*, 24(7):1011–1013, April 2008. PMID: 18204061.

[266] E Yang, P T Foteinou, K R King, M L Yarmush, and I P Androulakis. A novel non-overlapping bi-clustering algorithm for network generation using living cell array data. *Bioinformatics (Oxford, England)*, 23(17):2306–2313, September 2007. PMID: 17827207.

[267] W Yang, J R Diehl, and W E Roudebush. Comparison of the coding sequence of the platelet-activating factor receptor gene in three species. *DNA Sequence: The Journal of DNA Sequencing and Mapping*, 12(4):239–251, November 2001. PMID: 11916258.

[268] Yuko Yoshida, Yuko Makita, Naohiko Heida, Satomi Asano, Akihiro Matsushima, Manabu Ishii, Yoshiki Mochizuki, Hiroshi Masuya, Shigeharu Wakana, Norio Kobayashi, and Tetsuro Toyoda. 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, July 2009. PMID: 19468046.

[269] Shi Yu, Tillmann Falck, Anneleen Daemen, Léon-Charles Tranchevent, Johan A K Suykens, Bart De Moor, and Yves Moreau. L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, 11(1):309, June 2010. PMID: 20529363.

[270] Shi Yu, Léon-Charles Tranchevent, Bart De Moor, and Yves Moreau. Gene prioritization and clustering by multi-view text mining. *BMC Bioinformatics*, 11:28, 2010. PMID: 20074336.

[271] Shi Yu, Steven Van Vooren, Léon-Charles Tranchevent, Bart De Moor, and Yves Moreau. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics (Oxford, England)*, 24(16):i119–125, August 2008. PMID: 18689812.

[272] Wei Yu, Anja Wulf, Tiebin Liu, Muin J Khoury, and Marta Gwinn. Gene prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics*, 9:528, 2008. PMID: 19063745.

[273] Peng Yue, Eugene Melamud, and John Moult. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, 7:166, 2006. PMID: 16551372.

[274] E M Zdobnov and R Apweiler. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics (Oxford, England)*, 17(9):847–848, September 2001. PMID: 11590104.

[275] Peisen Zhang, Jinghui Zhang, Huitao Sheng, James J Russo, Brian Osborne, and Kenneth Buetow. Gene functional similarity search tool (GFSST). *BMC Bioinformatics*, 7:135, 2006. PMID: 16536867.

[276] Ning Zhong and Karl H Weisgraber. Understanding the basis for the association of apoE4 with alzheimer's disease: opening the door for therapeutic approaches. *Current Alzheimer Research*, 6(5):415–418, October 2009. PMID: 19874264.

[277] Jingde Zhu and Xuebiao Yao. Use of DNA methylation for cancer detection and molecular classification. *Journal of Biochemistry and Molecular Biology*, 40(2):135–141, March 2007. PMID: 17394761.

List of publications

Journal Papers

- 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", *Nature Biotechnology*, vol. 24, no. 5, May 2006, pp. 537-544.
- Heim T., **Tranchevent L.C.**, Carlon E., Barkema G.T., "Physical-chemistry-based analysis of affymetrix microarray data", *J. Phys. Chem. B.*, 2006 Nov 16;110(45):22786-95.
- De Bie T., **Tranchevent L.C.**, van Oeffelen L., Moreau Y., "Kernel-based data fusion for gene prioritization", *Bioinformatics*, vol. 23, no. 13, Jul. 2007, pp. i125-32.
- Thorrez L., Van Deun K., **Tranchevent L.C.**, Van Lommel L., Engelen K., Marchal K., Moreau Y., Van Mechelen I., Schuit F., "Using ribosomal protein genes as reference: a tale of caution", *PLoS One*, vol. 3, no. 3, Mar. 2008, pp. e1854.
- Yu S., Van Vooren S., **Tranchevent L.C.**, De Moor B., Moreau Y., "Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining", Bioinformatics, vol. 24, no. 16, Aug. 2008, pp. i119-125., Lirias number: 216972.
- **Tranchevent L.C.**, Barriot R., Yu S., Van Vooren S., Van Loo P., Coessens B., Aerts S., De Moor B., Moreau Y., "ENDEAVOUR update: a web resource for gene prioritization in multiple species", *Nucleic Acids Research*, Web Server issue, vol. 36, no. 1, Jun. 2008, pp. 377-384.
- Aerts S., Vilain S., Hu S., **Tranchevent L.C.**, Barriot R., Yan J., Moreau Y., Hassan B., Quan X., "Integrating Computational Biology and Forward Genetics in Drosophila", *PLoS Genetics*, vol. 5, no. 1, Jan. 2009, pp. 351. **(equally contributed author)**

- Thorrez L., **Tranchevent L.C.**, Chang H.J., Moreau Y., Schuit F., "Detection of novel 3'UTR extensions with 3' expression microarrays", *BMC Genomics*, 2010 Mar. 26;11:205. **(equally contributed author)**

- Barriot R., Breckpot J., Thienpont B., Brohee 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 Medicine*, vol. 2, no. 3, Mar. 2010, pp. 1-9.

- 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*, vol. 4, no. 5, May 2009, pp. e5526-.

- Yu S., **Tranchevent L.C.**, De Moor B., Moreau Y., "Gene prioritization and clustering by multi-view text mining", *BMC Bioinformatics*, doi:10.1186/1471-2105-11-28, vol. 11, no. 28, Jan. 2010, pp. 1-48.

- **Tranchevent L.C.**, Capdevila F.B., Nitsch D., De Moor B., De Causmaecker P., Moreau Y., "A guide to web tools to prioritize candidate genes", *Briefings in Bioinformatics*, vol. 11, no. 3, May 2010, pp. 1-11.

- Schuierer S., **Tranchevent L.C.**, Dengler U., Moreau Y., "Large-scale benchmark of Endeavour using MetaCore maps", *Bioinformatics*, 2010 Aug 1;26(15):1922-3. **(equally contributed author)**

- Yu S., Falck T., Daemen A., **Tranchevent L.C.**, Suykens J., De Moor B., Moreau Y., "L2-norm multiple kernel learning and its application to biomedical data fusion", *BMC Bioinformatics*, vol. 11, no. 309, Jun. 2010, pp. 1-53.

- Thienpont B., Zhang L., Postma A.V., Breckpot J., **Tranchevent L.C.**, Van Loo P., Mollgard K., Tommerup N., Bache I., Tumer Z., Van Engelen K., Menten B., Mortier G., Waggoner D., Gewillig M., Moreau Y., Devriendt K., Larsen L.A., "Haplo-insufficiency of TAB2 causes congenital heart defects in humans", *Am. J. Hum. Genet.*, 2010 Jun 11;86(6):839-49.

- Yu S., Liu X., **Tranchevent L.C.**, Glanzel W., Suykens J., De Moor B., Moreau Y., "Optimized data fusion for K-means Laplacian Clustering", Bioinformatics, vol. 27, no. 21, Jan. 2011, pp. 118-126.

- Yu S., **Tranchevent L.C.**, De Moor B., Moreau Y., "Kernel-based Data Fusion for Machine Learning", vol. 345 of Studies in Computational Intelligence, Springer, 2011, 200 p.

Manuscripts in preparation

- Thorrez L., **Tranchevent L.C.**, Lehnert S., Moreau Y., Schuit F., "Testis versus brain: 3'UTR size matters", Internal Report 08-113, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2008.
- Yu S., **Tranchevent L.C.**, Liu X.H., De Moor B., Moreau Y., "Integrating heterogeneous data sets for Clustering Analysis", Internal Report 08-200, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2008.
- **Tranchevent L.C.**, Leach S., Yu S., Moreau Y., "CrossEndeavour: Gene Prioritization using Multiple Species", Internal Report 09-141, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2009.
- Bonachela Capdevila F., **Tranchevent L.C.**, Moreau Y., de Causmaecker P., "Application of clustering in human gene prioritization using CLOPE and Endeavour", Internal Report 10-101, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2010.
- Yu S., Leach S., **Tranchevent L.C.**, De Moor B., Moreau Y., "Cross-species candidate gene prioritization with MerKator", Internal Report, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2010.
- Balikova I., **Tranchevent L.C.**, Moreau Y, Vermeesch J., "Designing an eye disorder specific microarray" Internal Report, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2010.
- Nitsch D., **Tranchevent L.C.**, Gonçalves J., Moreau Y., "PINTA - A web server for network-based gene prioritization from expression data", Internal Report 11-04, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2011.
- Nitsch D., **Tranchevent L.C.**, Bonachela Capdevila F., de Moor B., De Causmaecker P., Moreau Y., "CAGP : A Critical Assessment of Candidate Gene Prioritization", Internal Report 11-15, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2011. **(equally contributed author)**
- Bonachela Capdevila F., Nitsch D., **Tranchevent L.C.**, de Moor B., Moreau Y., De Causmaecker P., "A Pipeline through Gene Prioritization methods for identifying novel disease genes", Internal Report 11-16, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2011. **(equally contributed author)**
- Breckpot J., Thienpont B., **Tranchevent L.C.**, Gewillig M., Allegaert K., Vermeesch JR., Moreau Y., Devriendt K., "Congenital heart defects in a novel recurrent atypical 22q11.2 deletion syndrome harboring the genes MAPK1 and CRKL", Internal Report 11-17, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2011.

Curriculum vitae

Léon-Charles Tranchevent was born in Mayenne, France, in 1982. In year 2000, he obtained a high school diploma in sciences from the Lycée Raoul Vadepied of Evron, France. In year 2003, he obtained a Bachelor's degree in computer science from the Université Belle-Beille of Angers, France, with a thesis titled "Handwritten characters recognition via a matrix based and a vector based models". In year 2005, he completed a Master's programme in bioinformatics from the Université des Sciences et Technologies of Lille, France. His Master's thesis was realized at the Interdisciplinary Research Institute under the supervision of Prof. Enrico Carlon and was titled "Physical-chemistry based analysis of microarray data". In 2005, he joined the bioinformatics group of the Department of Electrical Engineering (ESAT-SCD) of the Katholieke Universiteit Leuven as a predoctoral student under the supervision of Prof. Yves Moreau. In 2006, he started a PhD program on "Gene prioritization through genomic data fusion" under the supervision of Prof. Yves Moreau.