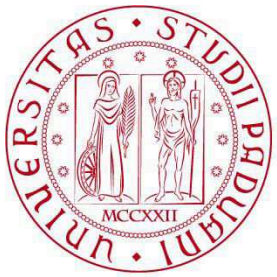


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova
Department of Biology

Ph.D. COURSE IN: BIOSCIENCES
CURRICULUM: GENETICS, GENOMICS AND BIOINFORMATICS
SERIES XXX

Prioritisation of candidate disease genes via multi-omics data integration

Coordinator: Prof. Ildikò Szabò
Supervisor: Dr. Rosella Tomanin
Co-Supervisor: Prof. Giorgio Valle

Ph.D. student: Guido Zampieri

Abstract

The uncovering of genes linked to human diseases is a pressing challenge in molecular biology, towards the full achievement of precision medicine. Next-generation technologies provide an unprecedented amount of biological information, but at the same time they unveil enormous numbers of candidate disease genes and pose novel challenges at multiple analytical levels. Multi-*omics* data integration is currently the principal strategy to *prioritise* candidate disease genes. In particular, kernel-based methods are a powerful resource for the integration of biological knowledge, but their use is often precluded by their limited scalability.

In this thesis, we propose a novel scalable kernel-based method for gene prioritisation which implements a novel multiple kernel learning approach, based on a semi-supervised perspective and on the optimisation of the margin distribution in binary problems. Our method is optimised to cope with strongly unbalanced settings where known disease genes are few and large scale predictions are required. Importantly, it is able to efficiently deal both with a large amount of candidate genes and with an arbitrary number of data sources. Through the simulation of real case studies, we show that our method outperforms a wide range of state-of-the-art methods and has enhanced scalability compared to existing kernel-based approaches for genomic data.

We apply the proposed method to investigate the potential role for disease gene prediction of metabolic rearrangements caused by genetic perturbations. To this end, we use constraint-based modelling of metabolism to generate gene-specific information at a genome scale, which is mined via machine learning. Moreover, we compare constraint-based modelling and our kernel-based method as alternative integration strategies for omics data such as transcriptional profiles. Experimental assessments across various cancers demonstrate that information on metabolic rewiring reconstructed *in silico* can be valuable to prioritise associated genes, although accuracy strongly depends on the cancer type. Despite these fluctuations, predictions achieved starting from metabolic modelling are largely complementary to those from gene expression or pathway annotations, highlighting the potential of this approach to identify novel genes involved in cancer.

Sommario

La scoperta dei geni legati alle malattie nell'uomo è una sfida pressante in biologia molecolare, in vista del pieno raggiungimento della medicina di precisione. Le tecnologie di nuova generazione forniscono una quantità di informazioni biologiche senza precedenti, ma allo stesso tempo rivelano numeri enormi di geni malattia candidati e pongono nuove sfide a molteplici livelli di analisi. L'integrazione di dati multi-*omici* è attualmente la strategia principale per *prioritizzare* geni malattia candidati. In particolare, i metodi basati su kernel sono una potente risorsa per l'integrazione della conoscenza biologica, tuttavia il loro utilizzo è spesso precluso dalla loro limitata scalabilità.

In questa tesi, proponiamo un nuovo metodo kernel scalabile per la prioritizzazione di geni, che applica un nuovo approccio di multiple kernel learning basato su una prospettiva semi-supervisionata e sull'ottimizzazione della distribuzione dei margini in problemi binari. Il nostro metodo è ottimizzato per fare fronte a condizioni fortemente sbilanciate in cui si disponga di pochi geni malattia noti e siano richieste predizioni su larga scala. Significativamente, è capace di gestire sia un gran numero di candidati sia un numero arbitrario di sorgenti di informazione. Attraverso la simulazione di casi studio reali, mostriamo che il nostro metodo supera in prestazioni un'ampia gamma di metodi allo stato dell'arte ed è dotato di migliore scalabilità rispetto a metodi kernel esistenti per dati genomici.

Applichiamo il metodo proposto per studiare il potenziale ruolo per la predizione di geni malattia dei riarrangiamenti metabolici causati da perturbazioni genetiche. A questo scopo, utilizziamo modelli del metabolismo basati su vincoli per generare informazione sui geni a scala genomica, che viene analizzata tramite apprendimento automatico. Inoltre, compariamo modelli basati su vincoli ed il nostro metodo basato su kernel come strategie di integrazione alternative per dati omici come profili trascrizionali. Valutazioni sperimentali su vari cancri dimostrano come i riarrangiamenti metabolici ricostruiti *in silico* possano essere utili per prioritizzare i geni associati, nonostante l'accuratezza dipenda fortemente dalla tipologia di cancro. Malgrado queste fluttuazioni, le predizioni basate su modelli metabolici sono largamente complentari a quelle basate su espressione genica o annotazioni di pathway, evidenziando il potenziale di questo approccio per identificare nuovi geni implicati nel cancro.

Contents

List of Figures

List of Tables

Acronyms

AUC Area Under the receiver-operating-characteristic Curve. 32–34, 37, 38, 44, 67–76, 83

CAKUT Congenital Abnormalities of Kidney and Urinary Tract. 43–45

CBM Constraint-Based Modelling. 4, 7, 8, 17, 18, 50, 52, 58, 61, 69, 70, 72–76, 78, 80, 81

CIF Conserved Intronic Fragments. 102–104

CPDB Consensus Path DataBase. 13, 44, 70, 72

epFBA euclidean parsimonious enzyme usage Flux Balance Analysis. 57, 64

FBA Flux Balance Analysis. 55, 56, 62, 64

GBA Guilt-By-Association. 8, 9, 59

GDAC Genome Data Analysis Center. 65, 96

GSMM Genome-Scale Metabolic Model. 7, 8, 17, 49, 50, 52–54, 57–59, 61, 62, 65, 66, 72, 79, 81

HPO Human Phenotype Ontology. 12, 39

KEGG Kyoto Encyclopedia of Genes and Genomes. 13, 33

KOMD Kernel-based Optimisation of Margin Distribution. 26

LEDK Laplacian Exponential Diffusion Kernel. 23, 34, 37

LOOCV Leave-One-Out Cross Validation. 32, 67–70

LSD Lysosomal Storage Disorder. 101–104

MDK Markov Diffusion Kernel. 23, 34–37, 44

MEDK Markov Exponential Diffusion Kernel. 23, 33, 34, 37

MEP Metabolic Expectation Propagation. 57–59, 67, 79

MKL Multiple Kernel Learning. 15–19, 24–26, 30, 35, 36, 70, 73–76, 79–81

NCI60 National Cancer Institute 60 human tumour cell line. 62, 66, 67, 70, 88

NGS Next-Generation Sequencing. 2, 16, 31, 44, 64, 101

OMIM Online Mendelian Inheritance in Man database. 1, 13, 32

PCC Pearson Correlation Coefficient. 33, 40, 62–64, 66

pFBA parsimonious enzyme usage Flux Balance Analysis. 56, 57

PPI Protein-Protein Interaction. 10, 11, 59

PU Positive-Unlabelled. 15, 28, 59, 77, 81

RHD Renal Hypo/Dysplasia. 43–45

RLK Regularized Laplacian Kernel. 24, 34, 37

SCC Spearman Correlation Coefficient. 36, 37, 69

Scuba SCalable UnBAlanced gene prioritization. 29–42, 44, 50, 59, 68, 70, 72, 77, 78, 81, 84, 85

SNP Single Nucleotide Polymorphism. 5, 103, 104

SVM Support Vector Machine. 15, 26, 77

TPR True Positive Rate. 37, 83

VEP Variant Effect Predictor. 103, 104

Mathematical symbols

$\mathbf{A}$ graph adjacency matrix. 22, 23

α diffusion parameter for LEDK, MEDK and RLK. 22–24, 33, 34

$\mathbf{b}$ metabolite intake/uptake vector. 51, 57, 58

$\mathbf{c}$ metabolite concentration vector. 50, 51

η vector of kernel coefficients. 25, 27

$\mathscr{G}$ set of genes. 27, 28, 32

Γ set of examples weight distribution. 26–28

γ examples weight distribution. 26–29

γ_- unlabelled examples weight distribution. 28

γ^{opt} optimal examples weight distribution. 26, 27, 29

Γ^+ set of positive examples weight distribution. 28, 29

γ_+ positive examples weight distribution. 28, 29

$\mathbf{I}$ identity matrix. 23, 24

$\mathbf{K}$ kernel matrix or Gram matrix. 21, 26

k kernel function. 20–22

$\mathbf{K}_{ED}$ Gram matrix of the exponential diffusion kernel. 22

$\mathbf{K}_{LED}$ Gram matrix of the Laplacian exponential diffusion kernel. 23

$\mathbf{K}_{MD}$ Gram matrix of the Markov diffusion kernel. 23, 83

$\mathbf{K}_{MED}$ Gram matrix of the Markov exponential diffusion kernel. 23

$\mathbf{K}^{-}$ sub-Gram matrix of unlabelled examples. 28

$\mathbf{K}^{opt}$ optimal Gram matrix. 24, 25, 27, 29, 30

$\mathbf{K}^{+}$ sub-Gram matrix of positive examples. 28, 29

$\mathbf{K}^{+-}$ sub-Gram matrix of positive-unlabelled example pairs. 28, 29

$\mathbf{K}_{RL}$ Gram matrix of the regularised laplacian kernel. 24, 83

$\mathbf{K}^{sum}$ sum of Gram matrices. 27–29

$\mathbf{L}$ Laplacian matrix. 23, 24

λ KOMD regularisation parameter. 26–28

λ_{-} Scuba regularisation parameter for unlabelled examples. 28, 35, 36

λ_{+} Scuba regularisation parameter for positive examples. 28, 29, 35, 44

μ gene set expression map parameter. 54, 61–64

$\mathscr{N}$ set of negative examples. 26

$\mathscr{P}$ set of positive genes. 26–28, 30, 32, 34

ϕ map from data points to a Hilbert space. 20, 21, 25

r Pearson correlation coefficient. 40, 62, 64

ρ Spearman correlation coefficient. 36, 37, 69

$\mathbf{S}$ stoichiometric matrix. 50–52, 55–57

s Scuba score. 29, 32

σ gaussian kernel parameter. 22, 44, 68, 70

t number of steps for MDK. 23, 24, 34, 44

$\mathscr{U}$ set of unlabelled genes. 28, 30, 32, 34, 35

$\mathbf{v}$ flux vector. 50, 51, 55–58

$\mathbf{v}_{lb}$ lower flux bound vector. 51, 54, 55

$\mathbf{v}_{ub}$ upper flux bound vector. 51, 54–57

$\mathbf{Y}$ label matrix. 26, 27

Chapter 1

Introduction

An avalanche of data is revolutionising molecular biology and has the potential to radically innovate disease diagnosis and treatment. In order to fully elucidate the genetic bases of human disease it is necessary to combine a variety of biological information within a complex systems framework. Computational methods allow aggregating knowledge and prioritising candidate genes of interest. However, their power remains limited as their development faces several challenges.

The identification of genes underlying human diseases is a major goal in current molecular genetics research. Despite the dramatic progresses of the last few decades, it is progressively emerging how little it is yet understood of the origin and evolution of genetically-based conditions. In the 1980s, only a handful of DNA loci were known to be related to disease phenotypes, as a result of massive investigative efforts. During the 1990s, positional cloning allowed mapping a vast portion of known Mendelian diseases to their causative genes [1, 2]. In positional cloning, disease genes are identified from their approximate chromosomal location, defined as tightly as possible. Most commonly, linkage studies lead to the identification of candidate chromosomal regions, where residing genes are tested. Alternative strategies are the use of chromosomal abnormalities and animal models. Nowadays opportunities for the diagnosis and the design of new therapies are progressively growing, thanks to several technological advances and the application of statistical or mathematical techniques.

However, despite the huge advances, even among Mendelian disorders much remains to be discovered. On December 21st 2016, the Online Mendelian Inheritance in Man database (OMIM) registered 4,908 Mendelian phenotypes of known molecular basis and 1,483 Mendelian phenotypes of unknown molecular origin [3]. Moreover, 1,677 more phenotypes are suspected to be Mendelian. Yet it is among poligenic and multifactorial

pathologies that the most remains to be elucidated, as the majority of them has been linked to only a few genetic loci [1, 2]. Traditional investigative techniques can be ineffective when complex non-additive effects arise from the action of multiple genes across different biological levels, especially when non-genetic factors further blur the picture. Evidence of complexity is emerging even among Mendelian disorders, stemming from a lack of correlation between genotype and the clinical phenotype [4]. Along with other factors, it is suggested that modifier genes may lie behind these observations, giving rise to the notion of oligogenic disease [5–8]. Moreover, once the putative disease genes have been identified, validating the true association may be challenging. Independent studies are often necessary to confirm the discovery. So far, clinically useful findings have been achieved in a few complex diseases.

In this context, the advent of high-throughput technologies radically changed the perspective towards the search for disease genes, along with the whole field of molecular biology. We start by describing the implications of this event for our comprehension of genetic disease.

1.1 An *omics* data perspective on human genetic disease

After the generation of the first complete genomes, the term *genomics* was coined to indicate the study of hereditary information as a whole. Following, a series of *omics* have born spanning several level of biological knowledge, most of them associated to emerging pieces of technology. Today, most widespread high-throughput technologies include DNA sequencing (genomics) [9], genome-wide RNA sequencing (transcriptomics) [10], methylation and histone modification data (epigenomics) [11, 12], protein mass spectrometry data (proteomics) [13] and metabolomics [14]. As technology moves forward, its cost decreases and growing wealth of data is generated. As an example, Figure 1.1 shows the average cost per genome since the time of first generation sequencing until the widespread commercialisation of most recent Next-Generation Sequencing (NGS) devices.

The "omics"suffix does not merely represent the entirety of a certain biological domain, but it has been more and more associated to a novel holistic understanding of biological systems. Inherited from the physics of complex systems, this vision lies on the observation that knowing the behaviour of components of a biological entity is not enough to understand its behaviour. The data revolution allowed thoroughly observing this fact for the first time, hence reinforcing the application of complex systems notions on the study of biological organisms. Terms like *systems biology*, *systems genetics* and *systems medicine* are today widespread, sanctioning the shift to a quantitative biology paradigm [15–20]. The underlying

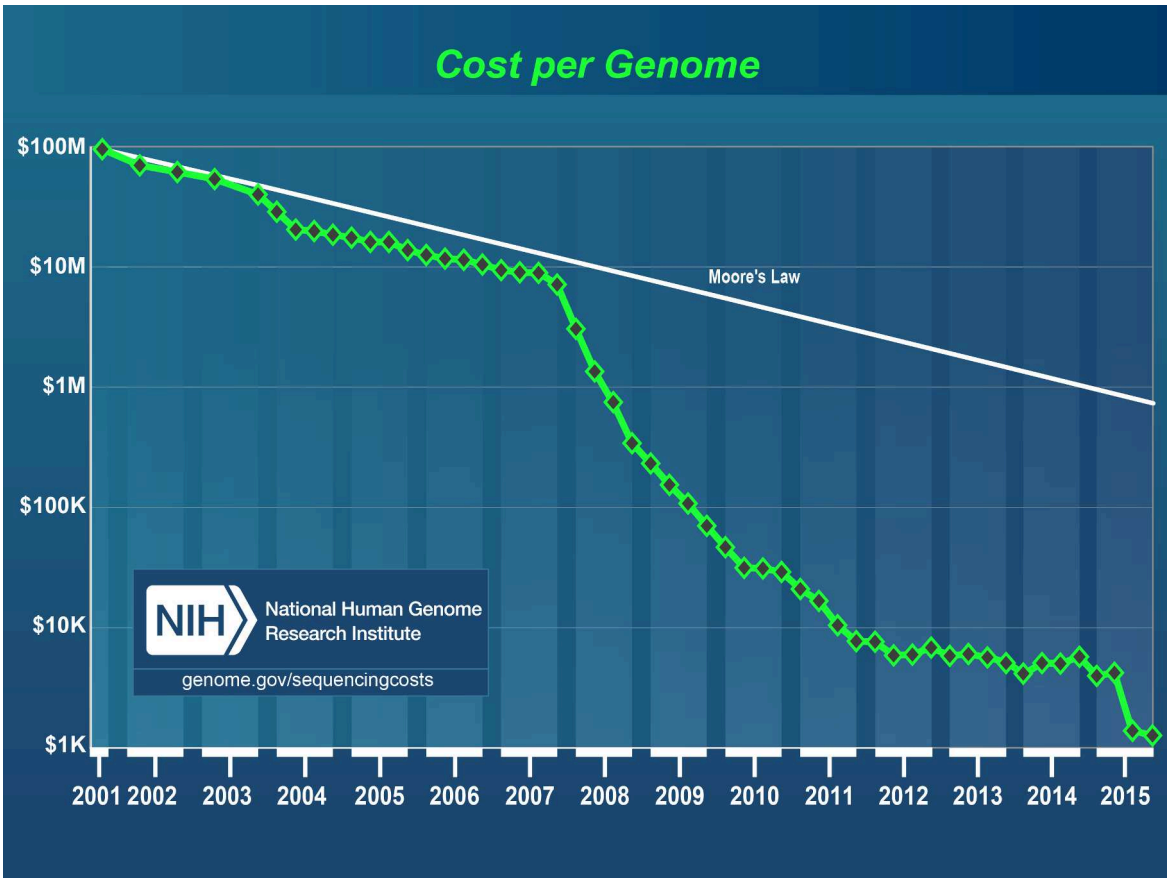


Figure 1.1 | Trend of whole-genome sequencing value. Average cost of whole-genome sequencing since the beginning of the millennium. After some years of strong development, we are now entering a phase of low and more stable value. The straight white line represents the theoretical cost as predicted by Moore's law, which is the empirical observation that the number of transistors in a dense integrated circuit doubles approximately every two years. This law is often taken as a goal by industry.

idea is that biological systems are governed by simple laws that lead to unexpected emerging phenomena.

On the level of human genetic disease, principal elements of complexity are:

- Gene pleiotropy: genes can have multiple functions, each of which is compromised by specific mutations.
- Gene interactions: diseases can be caused by the joined effect of several alterations in multiple genes.
- Interaction with epigenetic factors: epigenetic anomalies can alter or obfuscate the effect of genetic mutations.

- Interaction with metabolism and bio-molecular signalling: metabolic and signalling adjustments can compensate or modify the effect of genetic alterations.
- Interaction with environmental factors: infective agents, medications and diet can compensate or modify the effect of genetic alterations.
- Bio-molecular thresholds: genetic perturbations can have effect above determined functional thresholds.
- System dynamics: dynamical mechanisms can alter or obfuscate the effect of genetic alterations.

These elements ultimately manifest in the form of:

- Locus heterogeneity: same pathological phenotypes can result from different genetic alterations.
- Symptoms heterogeneity: same mutations can have different phenotypic effects depending on the context.

In order to translate the avalanche of accumulating information into knowledge, this has to be analysed and interpreted. It is more and more evident that our understanding of human disease remains limited because many studies have separately focused on the different biological levels, analysing single data types. In order to comprehend complex disorders aetiology, it is necessary to put together all pieces of the puzzle. In the following section we describe a general classification of current approaches for the integration of heterogeneous data aimed to the elucidation of genotype-phenotype relationships.

1.2 Data integration for unveiling interactions between genotype and human disorders

A single type of data offers a partial view on the complexity of the human biology and limits our understanding of it. Large-scale data integration approaches are conceived within a range of different mathematical and computational frameworks, traditionally divided between meta-dimensional and multi-staged methods [21]. Moreover, a third integrative framework has evolved in parallel, called Constraint-Based Modelling (CBM). In the next sections these three branches are described.

1.2.1 Multi-staged integration

As the name suggests, in multi-staged integration the data analysis is carried out through a sequence of consecutive steps. At each step, the information is filtered or processed in such a way to gradually let the relevant signal emerge. Multi-staged methods can be as diverse as the combination of target data sources, though they can be broadly described in terms of two main phases. A first phase involves the identification of relationships among the data sources via multiple cross-analyses. In a second and final phase, associations between data and phenotype are sought.

A common representative example of multi-staged approaches is the three-stage or triangle method. Here, Single Nucleotide Polymorphisms (SNPs) are first associated with the phenotype of interest and filtered on the basis of a genome-wide significance threshold. Second, significant SNPs are associated with another omic data type such as gene expression levels or methylation patterns to select genes harbouring those mutations. Finally, genes obtained from the previous step are tested for correlation with the phenotype. A different method consists in linking allele-specific expression with SNPs found in the corresponding genes and with cis-element variations or epigenetic modifications. In this case, allele-specific gene expression or methylation profiles are compared to identify differentially regulated alleles. Next, resulting genes are tested for correlation with a phenotype of interest.

This type of integration can be designed around the specific involved data sources, but it suffers from limitations both on a technical and on a biological assumptions level. Associations among data sets are in fact generally defined on arbitrary significance thresholds, which have to take into account also multiple hypothesis corrections. A number of true associations can thus be discarded. Furthermore, when phenotype is the result of a simultaneous complex interaction between multiple biological levels, a multi-staged approach will fail to reliably model it. This kind of methodology is thus indicated only for simple cases with nearly linear relationships between genotype and phenotype.

1.2.2 Meta-dimensional integration

Meta-dimensional methods simultaneously cross multiple data sources and can cope with variable inputs. They are broadly categorised into concatenation-based integration, transformation-based integration and model-based integration, whose schemas are displayed in Figure 1.2.

Concatenation-based integration fuses multiple data types together by concatenating data matrices into a single comprehensive matrix. Next, a model is generated starting from this combined matrix. An advantage of these approaches is the relative ease to apply statistical methods to any final data matrix. However, combining multiple matrices together can be

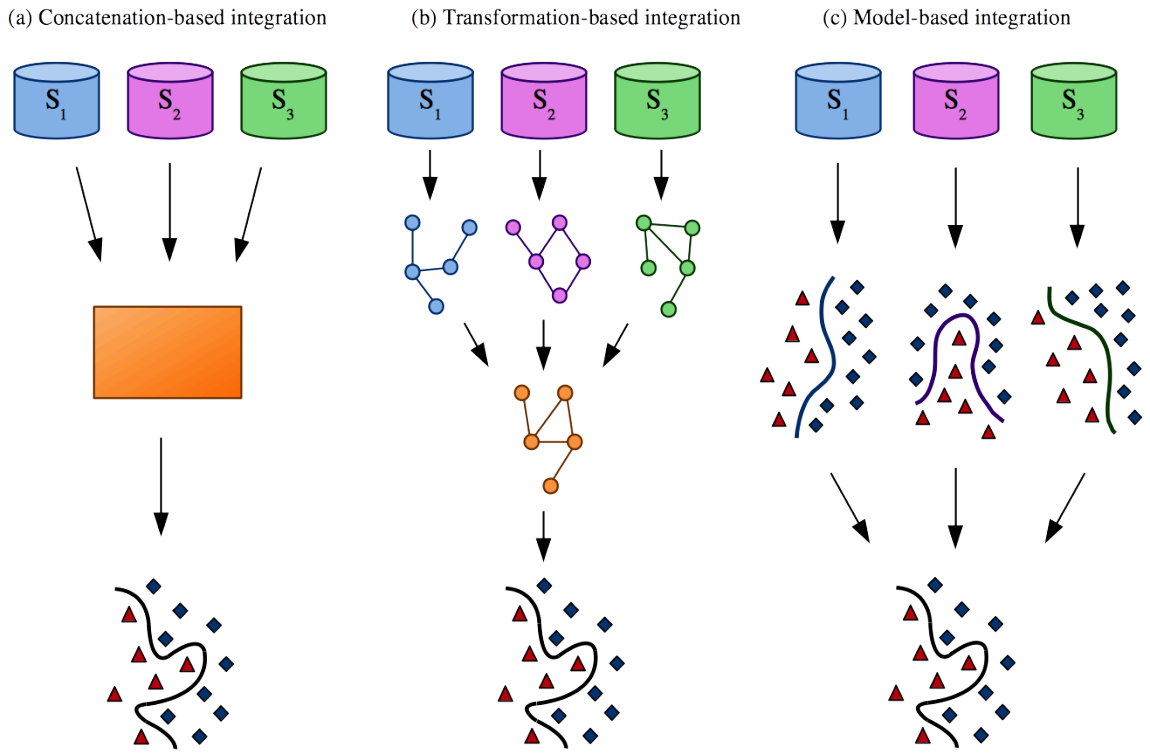


Figure 1.2 | Work-flow of meta-dimensional data integration approaches. Meta-dimensional integration methods can be broadly classified as: (a) concatenation-based integration, combining information at raw or processed data level; (b) transformation-based integration, combining information after conversion into a more abstract format; (c) model-based integration, generating models from data and next combining them to achieve the final model.

challenging, due to different scaling or inherent biases of each data type. Depending on the data considered, there are different degrees of difficulty in this phase. Moreover, sometimes it may be necessary a data reduction step if too many variables make analyses infeasible.

Transformation-based integration converts each dataset into an intermediate form such as a *graph* or a *kernel matrix*. The integration is performed at the level of transformed data, thus resulting in a integrative graph or matrix, which is used to obtain the model. These approaches have the advantage of preserving original data properties and they are able to combine virtually any data structure or format by converting it through appropriate transformations. A disadvantage is the difficulty to detect interactions among the different sources.

Model-based integration generates from each dataset an equal number of models, which are subsequently combined into a final model. This kind of integration can have an even larger flexibility as compared to other approaches. For instance, in patient-centred analyses

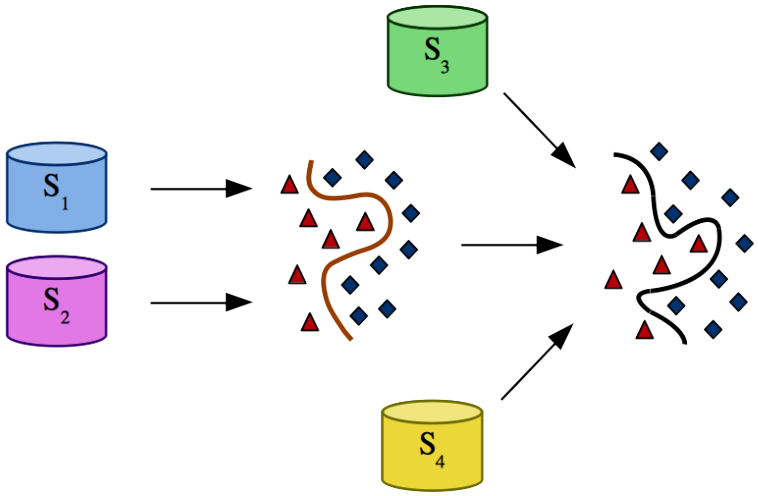


Figure 1.3 | Work-flow of constraint-based data integration. Metabolic models are generated from raw or processed data and available knowledge. In a successive step, additional data is mapped on the initial model to obtain a more refined model, specific for the cell type or the physiological condition of interest.

it is possible to combine models coming from different groups of patients. However, these approaches can miss interaction among different data types, just like transformation-based approaches. Furthermore, they are particularly sensitive to over-fitting, so they are indicated when the data pool is extremely heterogeneous.

1.2.3 Constraint-based integration

In the last few years, CBM has evolved in parallel as a novel type of data integration framework. From a formal point of view, it belongs to the field of systems modelling and revolves around metabolism, by means of Genome-Scale Metabolic Models (GSMMs). GSMMs include the whole set of biochemical reactions involved in cellular metabolism, along with their stoichiometry and directionality. Through CBM, it is possible to gain insights on cellular phenotypes in terms of reaction fluxes via a range of analyses [22].

As recently discussed, GSMMs can be the foundation for data integration [23–25]. In this case, the work-flow starts from raw data and knowledge on cellular physiology that are aggregated and converted into a GSMM. Later, additional omics data can be mapped onto the model, obtaining a new refined model. Mapped data can be transcriptional, proteomic and metabolomic profiles or information on splice isoforms or codon usage [26–28]. Depending

on the information introduced in the original model, a novel tissue-specific, cell-type specific or condition-specific model is created. General purpose models are used as scaffolds for such model-driven data integration, as the recent human cell models Recon 2.2 and HMR 2 [29, 30].

CBM-based data integration and analysis also have their limitations [31]. First, there are practical limitations in validating predictions, either because of technical issues in measuring the metabolites in question or difficulty of accessing the patient materials. Second, since GSMMs focus on metabolic enzyme-coding genes, reactions and pathways, GSMMs cannot be used to study genes involved in signal transduction or other functional domains. Third, GSMMs are based on the assumption of steady-state conditions and do not contain detailed reaction kinetics, so predicted flux distributions are often approximations of real reaction rates.

In Chapter 4 a more detailed description of approaches to integrate transcriptional data with metabolic models is provided.

1.3 Candidate gene prioritisation

Independently of the type of disease, the search of causative genes usually concerns a large number of suspects. It is therefore necessary to recognise the most promising candidates to submit to additional investigations, as experimental procedures are often expensive and time consuming. Gene prioritisation is the task of ordering genes from the most promising to the least. In traditional genotype-phenotype mapping approaches - as well as in genome-wide association studies - the first step is the identification of the genomic region(s) wherein the genes of interest lie. Once the candidate region is identified, the genes there residing are prioritised and finally analysed for the presence of possible causative mutations [1]. More recently, in new generation sequencing studies this process is inverted as the first step is the identification of mutations, followed by prioritisation and final validation [32].

Candidate gene prioritisation is therefore a necessary step in various sorts of investigations: from linkage analyses to genome-wide association studies and so on. This task is strictly related to gene selection, where the output is binary acceptance/rejection of candidate genes. Indeed, many methods can perform both prioritisation and selection. For instance, if a method assigns a score to each gene, it is easy to sort genes based on the scores or set a threshold to select only a fraction of them. However, prioritisation can leave more freedom to biologists, letting them decide how deeply scan the sorted candidate list.

Prioritisation criteria are usually based on functional relationships, co-expression and other clues linking genes together. In general, they follow the Guilt-By-Association (GBA)

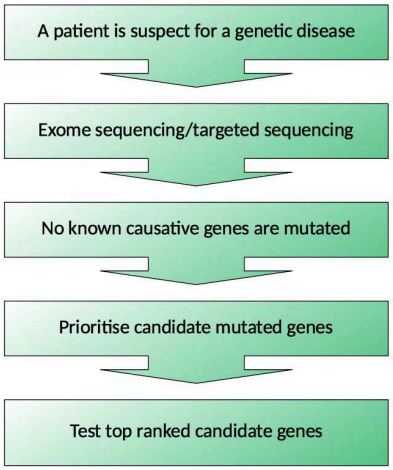


Figure 1.4 | Disease gene identification work-flow. Rationale of the procedure to identify novel gene-disease associations in modern next generation sequencing studies in clinical settings.

principle, i.e. disease genes are sought by looking for similarities to genes already associated to the pathology of interest [1, 33, 34]. Although GBA is the main strategy and has proven successful in a number of cases, it has the major limitation of assuming static relationships - a strongly simplistic assumption. In the attempt to overcome this, it has been proposed to associate genes by focusing on the perturbations in gene networks associated to disease states via a guilt-by-rewiring approach [35]. In another study, disease genes are correctly identified starting from changes in protein networks controllability upon mutations [36].

It is important to remark that prioritisation strategies can be targeted to the optimisation of virtually any kind of experimental investigation. Given the strong interest in the complete elucidation of genetic disorders aetiology, they are commonly employed in the prioritisation of candidate disease genes. Throughout this thesis, only this case will be considered, tacitly implying that the same considerations can be applied to any situation where a prioritisation is required.

The individuation of genes involved in a certain phenotype is usually not enough to fully comprehend its molecular roots, as it depends on the specific alterations in that gene. It is thus necessary to understand also how genomic variants can impact gene functions. Such a task can be formulated as variant prioritisation, where large sets of variants are ranked on the basis of various biological relationships - analogously to the gene-centred setting [37–39]. This problem shares some difficulties with gene prioritisation, namely the positive-unlabelled and imbalanced set-up - as will be discussed in Section 2.5 - but it is out of the scope of this thesis [38].

1.3.1 Characteristics of disease genes

Disease genes tend to present some peculiar characteristics: statistically speaking, they have longer genomic and amino acid sequence, span a broader phylogenetic extent and present specific conservation and paralogy profiles as compared with all human proteins [40, 41]. Even in relation to house-keeping genes, they show different evolutionary conservation rates, DNA coding lengths and gene functions [42]. A phylogenetic analysis has revealed that many disease genes originated during the early evolution of metazoa, while genes specific to the mammalian lineage are strongly under-represented [43]. Moreover, disease genes evolve at higher synonymous substitution rates with higher nonsynonymous/synonymous substitution rate ratios and are expressed in a narrower range of tissues [44–46]

Perhaps more influential to gene prioritisation research are observations on functional properties in the context of biological networks. Development of graph theory enforced a network-centred view of biological systems. Indeed, proteins with common or similar functions are seen to cluster together in physical protein interaction networks. In this context, disease is interpreted as disruption of these and other bio-molecular networks and genes associated with the same disease are believed to be preferentially interconnected [47–50]. In light of these findings, networks represent a major tool for the search of disease genes, as is described in the next section.

1.4 Omics data for gene prioritisation

As described in the previous section, disease genes often possess specific distinctive features. In light of these observations, several types of biological evidence can be used to recognise them. Depending on the disorder of interest and on the available information, most appropriate data sources can be chosen *ad-hoc* in order to orient the prioritisation. In the following we outline the principal categories, presenting their assumptions and limitations.

Interactomes The term *interactome* refers to the global set of molecular interactions in a cell such as Protein-Protein Interactions (PPIs), protein-nucleic acid interactions, protein-metabolites interactions or interactions between post-translational modifiers and their targets. In particular, PPIs were the first to be extensively studied. Even considering one splice variant per coding gene, at least 20,000 proteins must exist implying a huge amount of putative interactions [51–53]. Since 1994, the number of identified high-confidence binary PPIs has indeed grown roughly linearly up to ∼11,000 interactions in 2013 [54]. They have been aggregated, together with other ∼14,000 newly generated ones, into the widest and most rigorously verified protein

interactome to date [54]. More recently, large-scale efforts have allowed widening the boundary of previous interaction maps through affinity purification–mass spectrometry methodology, resulting in BioPlex and BioPlex 2.0 [55, 56], while novel gene editing techniques promise to push forward interactome exploration in the future [57]. Protein interactions can also be predicted starting from a variety of biological hints such as sequence homology, gene co-expression, phylogenetic profiles and three-dimensional structure [58]. Nevertheless, PPIs represent only a portion of biomolecular interactions involved in pathological mechanisms. Since recently, first maps are available also for interactions between messenger RNA, micro RNA and long non-coding RNA molecules [59–62]. Finally, another kind of interactome is represented by genetic interactions (Figure 1.5).

Interactomes are one of the most widely utilised data source to prioritise genes. However, there are still major limitations arising from experimental issues or poorly investigated aspects. The main drawback is the bias toward well-characterised genes or gene products, that make up the dominant portion of current maps [63]. This is due to the fact that most known interactions have been identified in small-scale studies, because of limited resources and technical challenges. Secondly, reconstructed networks do not take into account that physical interactions are dependent on tissue, cell type and physiological conditions [64]. Quantitative and dynamic features, such as protein expression levels and interaction strength, are not yet incorporated [63].

Gene expression Genome-wide experimental data like gene expression quantification represent an alternative information source with coverage on the full exome or genome. Inference on putative disease genes is usually performed in terms of differential expression between healthy and affected samples or of co-expression across different tissues and conditions [66]. This can be done through the construction of networks where distance among genes is weighted by their differential expression or co-expression [67]. Alternatively, this data type is analysed by means of statistical vector correlation measures. Gene expression can also provide information on specific tissues.

However, gene expression can be a poor indicator of the actual phenotype. In fact, some studies show that gene expression alone fails in enhancing classification of patient tissues [68]. Moreover, there is evidence suggesting that co-expression can often be due to genomic proximity rather than to functional links [69].

Functional annotations Biological functions are inherently connected to disease, as the latter arises when some function is compromised at a molecular level. Determining the function of a gene therefore corresponds to defining its associated disorders. Several

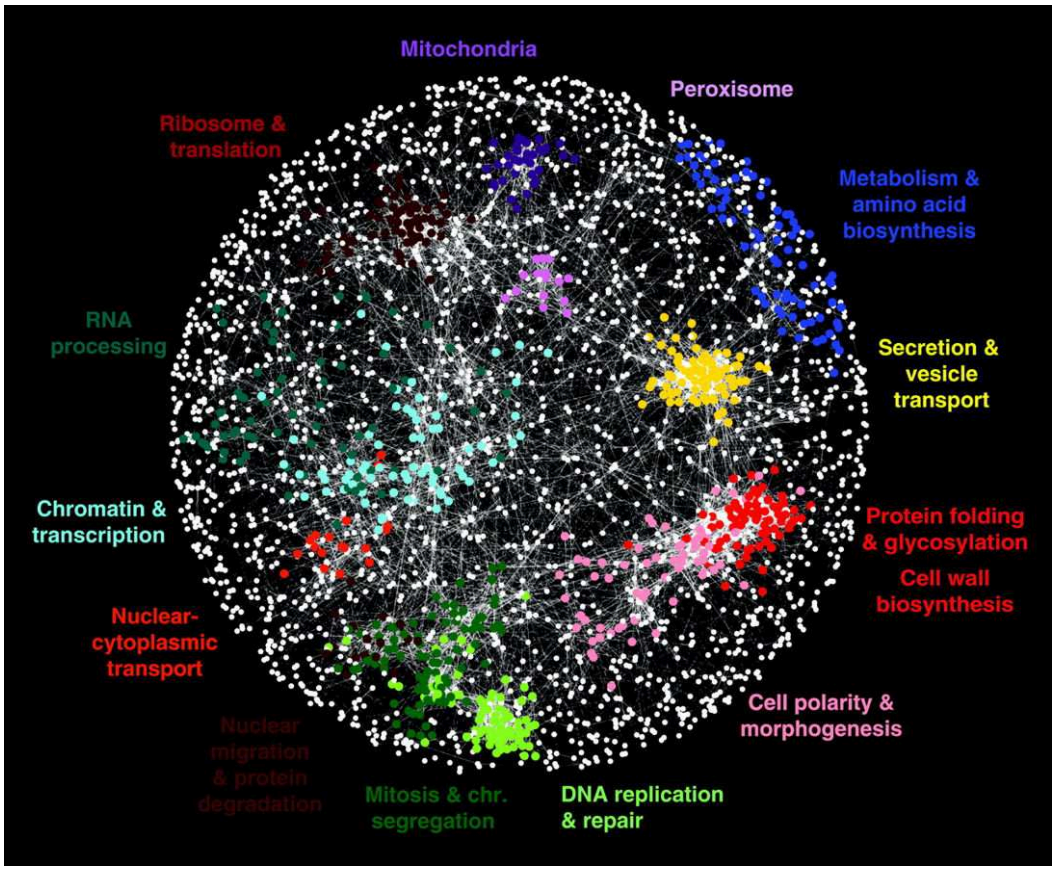


Figure 1.5 | Genetic network of a cell. Network among genes with similar genetic interaction profiles in Saccharomyces cerevisiae. Clusters of functionally related genes are highlighted by different colours (Source: Costanzo *et al* [65]).

types of evidence can provide information on the function of genes. The Gene Ontology provides structured annotations on three level of knowledge: biological processes, molecular functions and cellular localisation [70]. Other resources are the Human Phenotype Ontology (HPO) and the Mammalian Phenotype Ontology [71, 72].

These annotations provide an abstract and hierarchical organisation of knowledge which unifies and sorts a vast amount of experimental data. Limitations are that they are static and often suffer from high rates of incorrect annotations.

Pathway composition Biological pathways are chains of biochemical reactions that determine the production of a certain molecule or a well-defined change in a cell. They represent higher order biological functions, grouping together interactions and processes among molecules. Pathways are broadly classified based on their role in metabolism, genetic regulation and transmission of signals. Perhaps the most popular

pathway resource is Kyoto Encyclopedia of Genes and Genomes (KEGG), followed by Reactome [73, 74]. The most complete is Consensus Path DataBase (CPDB), which gathers information from 32 external repositories including KEGG and Reactome [75].

Gene products that participate in a certain pathway share functional complicity and by extension also their genes. The rationale for pathway-driven gene prioritisation is therefore to search for genes involved in the same pathways of known disease genes. A limitation for these methods is usually to not take into account the actual role of genes in the pathways but merely their presence.

Text mining The compendium of scientific articles in PubMed and similar resources provide a huge amount of knowledge that can be analysed through automatic language processing techniques to search for hidden associations between genes and disease [76]. This kind of analysis mimics the literature review usually performed by researchers, but at rates orders of magnitude faster. In the same way, other textual resources like OMIM can be mined [3]. Phenotypes without referenced molecular basis can be characterised by other means and compared with each other [77]. For instance, through text mining clustering of genes in terms of associated phenotype descriptions was found to correlate with their functional annotations [78].

A problem for this kind of approach is represented by a lack of consistent concept representation, as the same concepts can be formulated and phrased through various synonyms and aliases. Controlled vocabularies were generated to contrast this issue, providing standard references [79, 76]. Moreover, extracted information tends to be biased toward well studied genes, as those not covered in the literature lack meaningful information.

Sequence information Specific properties of DNA or gene product sequence can contain useful information. First, sequence may reflect functional aspects. Moreover, disease genes are statistically characterised by a higher length of coding region and other peculiar features, as described in a previous section.

Integrated gene networks Some resources gather gene relationships on various levels and merge them into an integrated gene network. These networks can be analysed through the same principles and techniques used for physical and genetic interactomes. For instance, STRING puts together functional associations across seven types of experimental or *in silico* evidence [80].

1.5 Computational gene prioritisation

In the last few years, computational techniques have been developed to aid researchers prioritise candidate genes, applying both statistics and machine learning [81–84]. As discussed in Section 1.1, a huge amount of data is in fact available for this kind of investigations. In particular, computational methods are essential for multi-*omics* data integration [21, 85]. Just like in related tasks, clues are often embedded in different data sources and only their combination leads to the emergence of informative patterns. Furthermore, incompleteness and noise of the single sources can be overcome by inference across multiple levels of knowledge.

Gene prioritisation methods can be broadly classified on four main levels: mathematical and computational approach, required input, type of biological evidence employed and scope of application. In the following a general description of these aspects is provided.

Computational approach Machine learning is a field of study that intersects mathematics, statistics and computer science and aims to the identification and analysis of patterns in the data. During the last two decades, this field has progressively invaded genetics and genomics as a novel investigation tool on a broad spectrum of problems [86, 87]. Machine learning can be broadly divided in supervised and unsupervised methods. The former are typically used to discriminate between two or more sets of items, such as genes. The latter are used to group items without any response variable. Most computational gene prioritisation methods are grounded on a supervised or semi-supervised machine learning approach (Chapter 2). Other methods use a purely statistical approach.

Input Most methods require some input data about the disease of interest, typically its known associated genes but also phenotypic or biochemical information. They are therefore disease-centred. On the other hand, some methods provide an overall evaluation or probability on the involvement of candidates in any disorder. This approach is grounded on the observation that disease genes tend to have peculiar characteristics, as described in Section 1.3.1. This second category of methods may be useful when it is not possible to have a meaningful biological reference for disease-centred approaches.

Type of evidence As discussed in the previous Section, various information sources can be used for prioritisation. They include physical and genetic interactions, gene expression, pathway composition, functional annotations, text mining, phenotype relationships, sequence data and regulatory information. Several combinations of data types can potentially be useful, so methods have been developed focusing on different specific

data types. Others are more comprehensive and are build to integrate virtually any number and type of data source.

Scope of application Some methods have been developed specifically for some classes of disorders, such as metabolic diseases or cancers. Conversely, other methods are generic, in the sense that they can be used without limitations on the disease type.

1.5.1 Methods and tools for gene prioritisation

One of the first methods to appear was Endeavour, which uses a statistical algorithm based on order statistics [88]. Distinct rankings are generated from multiple data sources and are fused together into a global ranking. Still today, it is one of the most popular methods, as it is available as web tool and it is considered a strong benchmark.

Most of other methods rely on machine learning algorithms. In particular, a class of algorithms is based on *kernels*, which are mathematical transformations that permit to estimate the similarity among items (in our case genes) taking into account complex data relations [89]. Importantly, kernels provide a universal encoding for any kind of knowledge representation, e.g. vectors, trees or graphs. When data integration is required, a Multiple Kernel Learning (MKL) strategy allows a data-driven weighting/selection of meaningful information [90]. The goal of MKL is indeed to learn optimal kernel combinations starting from a set of predefined kernels obtained by various data sources. Through MKL the issue of combining different data types is then solved by converting each dataset in a kernel matrix. Given the capacity of integrating virtually any type of data and the variety of existing kernels, this class of methods is particularly flexible and promising.

Numerous Multiple Kernel Learning (MKL) approaches have been proposed for the integration of genomic data [91, 92] and some of them have been applied to gene prioritisation [93–96]. De Bie *et al* formulated the problem as a one-class Support Vector Machine (SVM) optimisation task [93], while Mordelet and Vert tackled it through a biased SVM in a Positive-Unlabelled (PU) framework [95, 97]. Recently, Zakeri *et al* proposed an approach for learning non-linear log-Euclidean kernel combinations, showing that it can more effectively detect complementary biological information compared to linear combinations-based approaches [96]. However, as highlighted in a recent work by Wang *et al* [91], current methods share two limitations: high computational costs - given by a (at least) quadratic complexity in the number of training examples - and the difficulty to predefine optimal kernel functions to be fed to the MKL machine.

Besides kernel-based methods, a wide range of approaches have been proposed in the last two decades. Most of them are progressively converging towards the integration of

heterogeneous multi-omics data, and particularly of network data [98–101]. Moreover, various gene prioritisation web tools are currently available [102].

1.6 Research motivations and objectives

The present thesis is framed within BioInfoGen, a *strategic project* of the University of Padova started at the beginning of 2014 with the objective of developing technology and expertise in bioinformatics applied to personal genomics. This field of research requires multi-disciplinary expertise, provided by the joint contribution of different research groups from the Department of Biology, the Department of Mathematics, the Department of Women's and Children's Health and the Department of Medicine.

Units of the Department of Medicine and the Department of Women's and Children's Health are routinely involved in the diagnosis of patients via NGS and analysis of mutations, exploiting a sequencing facility of the Unity at the department of Biology. As in typical diagnostic screens, they often come across long lists of candidate genes and need to prioritise them. Upon an initial evaluation of available web tools, we decided to investigate novel prioritisation strategies.

Given the nature and the heterogeneity of the data that need to be integrated in order to achieve effective prioritisations, kernel methods appear as both powerful and flexible tools. Indeed, they can detect complex patterns robustly and efficiently from a limited data sample and at the same time facilitate the data integration through the kernel encoding. However, their application to the gene prioritisation problem has not been extensively investigated, consequently we decided to employ kernel methods as methodological basis.

As discussed in Section 1.5.1, a major limitation of current kernel-based methods for gene prioritisation is the limited scalability to large sets of genes and volumes of data. The BioInfoGen mathematics research unit has solid expertise in this field and in particular on the development of a scalable MKL algorithm [103]. Such a tool has however never applied to gene prioritisation nor possesses all desired scalability properties.

In light of these premises, the first objective of this work is to build a novel MKL method for gene prioritisation. Primary desired properties are accuracy in predictions and scalability to large lists of candidates.

We validated the proposed method on the retrieval of known gene-disease associations relative to a broad spectrum of disorders, in two main evaluation settings. First, we benchmarked it by means of a standard procedure called *cross-validation*. This analysis permits to estimate the average prediction error by repeated train-and-test assessments. Secondly, we

focused on a validation setting introduced specifically for gene prioritisation, where the goal is to estimate the accuracy in real circumstances when looking for unknown gene-disease associations. Here evaluation was restricted to predictions for associations discovered later than the last update of the data sources.

During the first part of the project, we observed from literature and experimental analyses that each data source has its intrinsic biases, as discussed in Section 1.4. This verification highlights the necessity for strategies to reduce them, either by developing new computational methods or by employing different complementary data. In particular, metabolic information is seldom considered, also due to technological limitations that still hamper large-scale metabolite profiling. At the same time, metabolism is emerging as a fundamental aspect in several type of diseases, such as cancers and neurological disorders, besides strictly metabolic pathologies. As presented in Section 1.2.3, CBM can be used to estimate genome-wide metabolic states even integrating disorder-specific experimental data and to study the interaction between genotype and metabolism. In this context, genetic perturbations can be analysed in terms of their down-stream metabolic rewiring. However, value of combining CBM and machine learning for gene prioritisation has never been tested before. As an additional point, the BioInfoGen unit of the Department of Women's and Children's Health focuses its research on metabolic disorders and has therefore strong interest in new tools for this kind of pathologies.

For these reasons, the second objective of this work is to develop and test a novel approach for gene prioritisation that combines CBM with MKL, assessing whether integrating *in silico* metabolic flux information can result in improved predictions. Moreover, from an operational point of view GSMMs can be used both as data sources and as data integration scaffolds, alternative to MKL. Therefore, we also aim to compare these two frameworks on the basis of their final prioritisation accuracy.

In order to integrate MKL and metabolic modelling, we developed a pipeline to (i) generate *in silico* flux profiles estimating the impact of each prioritised gene in the disorder; (ii) analyse obtained flux profiles via the learning algorithm developed in the first part of the thesis. We validated this approach focusing on cancer genes, independently considering the prioritisation of oncogenes and tumour suppressor genes. Tests were performed in the form of cross-validation, evaluating the complementarity of predictions in comparison to those obtained from other data sources. Finally, we performed analogous analyses to compare effectiveness of MKL and CBM as alternative omics data integration strategies.

1.7 Thesis outline

This thesis is divided in two parts, where we investigate two different aspects of data integration for gene prioritisation.
In the first part, we focus on the methodology for automatic gene prioritisation. In particular, we address the need for efficient methods that cope with the increasing volume and diversity of biological data. The first part of Chapter 2 introduces to the mathematical formulation of kernel methods and describes the MKL approach. The second part presents a novel scalable MKL method for heterogeneous omics data integration and gene prioritisation. In Chapter 3, experimental results of the proposed method are presented, along with a comparison with existing methods and tools.
In the second part, we focus on the integration between CBM and MKL aimed to prioritise candidate genes involved in metabolism. In Chapter 4 we first describe in detail the theory regarding CBM that allows investigating genetic perturbations at a metabolic flux level. In the final part, we advance a methodology for implementing these concepts in gene prioritisation. In Chapter 5 we present the application of the proposed approach to the prioritisation of tumour suppressor genes and oncogenes.
Finally, in Chapter 6 contributions and limitations of the present study are discussed and future research directions are delineated.

Chapter 2

Data integration via class-unbalanced multiple kernel learning

Kernel methods are a class of machine learning approaches that allow modelling non-linear data relationships. Among them, multiple kernel learning is a flexible strategy to integrate heterogeneous data, but it usually requires strong computational effort. We present here a novel scalable kernel-based method to effectively prioritise candidate genes on a genome-scale.

Kernel methods are advanced machine learning techniques, known for their robustness and stability in detecting stable patterns in the data. They are typically composed of two main steps. Initially, data is mapped to a higher dimensional space via a shortcut called *kernel function*. Due to the higher dimensionality, complex non-linear relationships in the original space appear linear in the new space. This operation takes into account the particular data type and domain knowledge on the patterns that one seeks to discover. Next, a pattern analysis algorithm is used to detect hidden relationships uncovered by the kernel function. Such an algorithm does not need to be suited for non-linear patterns - as these are already captured by the kernel - but it can be a simple linear algorithm. From a computational point of view, kernel methods have two important advantages. First, they allow identifying complex non-linear patterns at low computational cost in space and time. Second, despite the complexity that kernel functions can assume, learning algorithms are generally based on the optimisation of convex functions and thus are not threatened by local minima.

In this Chapter, we introduce the reader to Multiple Kernel Learning (MKL), a class of kernel methods that are particularly apt to integrating heterogeneous data. Following, we present a novel MKL algorithm, specifically designed for candidate gene prioritisation.

2.1 Foundations of kernel methods

Kernels can be informally seen as similarity measures between pairs of data examples. Consider the common case where each example is described in terms of a vector of real numbers. For instance, where examples are genes, they could be associated to a vector of expression values in different conditions. Mathematically, a similarity between two real vectors can be provided by their inner product. Indeed, if the vectors are parallel their inner product has maximum value and is equal to the product of their Euclidean norm. If the vectors are orthogonal, their inner product is null.

Kernel are similarities defined by inner products between vectors in a Hilbert space $\mathscr{H}$. A Hilbert space is an infinite-dimensional space having certain properties that guarantee that it is isomorphic to $\mathbb{R}^n$ for some finite n [89]. Denoting with $\mathscr{X}$ the global set of data instances, a kernel function k on $\mathscr{X} \times \mathscr{X}$ is formally defined as:

$$\begin{aligned} k &: \mathscr{X} \times \mathscr{X} \longrightarrow \mathbb{R} \\ k(x_1, x_2) &= \langle \phi(x_1), \phi(x_2) \rangle , \end{aligned} \tag{2.1}$$

where $x_1, x_2 \in \mathscr{X}$ and $\phi : \mathscr{X} \longrightarrow \mathscr{H}$ is a mapping from $\mathscr{X}$ to a feature space $\mathscr{H}$. In other words, via kernel functions one is able to obtain the dot product among data instances in $\mathscr{H}$ without calculating an infinite number of coordinates. This technique is known in the machine learning community as the *kernel trick*. The application of a kernel function to data

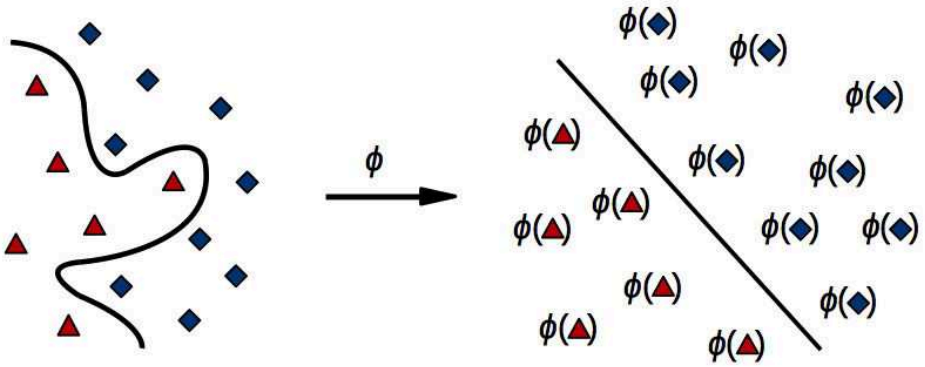


Figure 2.1 | Mapping to a Hilbert space. Input data can be embedded into an higher dimensional feature space via a map ϕ. As an effect of the higher dimensionality, non-linear patterns become linear.

generates a so called **kernel matrix** or **Gram matrix** $\mathbf{K}$. Such matrix has entries

$$\mathbf{K}_{ij} = \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j) \,. \tag{2.2}$$

To be a kernel function, k needs to be:

i. symmetric, i.e. $k(x_i, x_j) = k(x_j, x_i)$ and $\mathbf{K}_{ij} = \mathbf{K}_{ji}$;

ii. semi-definite, i.e. the kernel matrix $\mathbf{K}$ has all eigenvalues ≥ 0 .

Once the kernel matrix is calculated for any given dataset, the learning algorithm can extract predictive models from it. Kernel methods thus follow a simple modular work-flow shown in Figure 2.2. As a consequence, it is fundamental that $\mathbf{K}$ encodes all relevant information to the task. In the following section, we present some important types of kernel functions that can be useful to gene prioritisation.

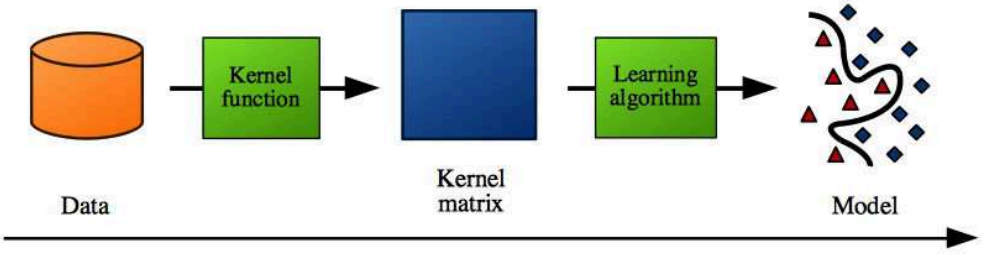


Figure 2.2 | Work-flow of kernel methods. The first step is the creation of a Gram matrix containing similarities among data instances through an appropriate kernel function; next, a predictive model is obtained from the kernel matrix via an algorithm for linear pattern analysis.

2.2 Types of kernels

Kernels can be used to define similarities starting from various data types, like strings or graphs. Some kernels can be analytically computed, while others need recursions or sampling procedures. In omics data for gene prioritisation, described in Section 1.4, genes are mainly represented in the form of real numbers vectors - e.g. gene expression profiles, annotations - or of network nodes - e.g. interactomes. In this work, we therefore focus on kernels defined over these kind of data.

2.2.1 Kernels on vectors

One of the most popular kernels defined on vectors is the **Gaussian kernel**. Given two data instances $\mathbf{x_1}$ and $\mathbf{x_2}$, their Gaussian kernel is defined as:

$$k(\mathbf{x_1}, \mathbf{x_2}) = e^{-\frac{\|\mathbf{x_1}-\mathbf{x_2}\|^2}{2\sigma^2}}. \tag{2.3}$$

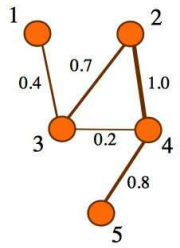
This function decreases exponentially as $\mathbf{x_1}$ and $\mathbf{x_2}$ get more distant from each other in the input space and when $\mathbf{x_1} = \mathbf{x_2}$ their similarity is maximal and equals 1. Here σ is a parameter that controls the decay velocity of the exponential. Small values of σ make the kernel matrix approaching the identity matrix, hence leading to data over-fitting and precluding the identification of any meaningful pattern. Conversely, large values lead k to come close to a constant function, making it infeasible to learn any non-trivial predictor. In other words, data examples are projected to an infinite-dimension feature space, where the weight of the features is controlled by σ. The higher is the parameter value and the more higher-order features dominate over lower-order ones.

2.2.2 Kernels on network nodes

A graph $G = (V, E)$ is a structure consisting of a **vertex** (or node) set $V = \{v_1, \ldots, v_N\}$ and a set $E = \{(v_i, v_j) | v_i, v_j \in V)\}$ where each node pair is called **edge** and represents a connection between the nodes. An efficient representation for graphs is the **adjacency matrix** $\mathbf{A}$, defined as the $N \times N$ matrix whose elements A_{ij} correspond to the number of links between two nodes i and j. In the scope of bio-molecular networks, we can restrict to the case where edges are undirected and weighted, i.e. they represent unordered pairs of nodes and have associated a real scalar number representing the connection strength. Thus $\mathbf{A}$ is symmetric and it contains real values as depicted in the simple example of Figure 2.3.

A graph node kernel aims at defining a similarity between any pair of nodes in a graph. The rationale is to characterise its nodes in terms of their direct or indirect connections. Numerous graph node kernels have been introduced, particularly in the field of recommender systems [104]. The most popular is the exponential diffusion kernel, which is based on the mathematical description of heat diffusion [105]. Its key idea is to assume a given amount of heat on each node and let it *diffuse* through the edges. The similarity between two nodes v_i, v_j is then measured as the amount of heat starting from v_i and reaching v_j over an infinite time interval. Formally, we have:

$$\mathbf{K}_{ED} = \sum_{t=0}^{\infty} \frac{\alpha^t \mathbf{A}^t}{t!} = e^{\alpha \mathbf{A}}, \tag{2.4}$$



$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.7 & 1 & 0 \\ 0.4 & 0.7 & 0 & 0.2 & 0 \\ 0 & 1 & 0.2 & 0 & 0.8 \\ 0 & 0 & 0 & 0.8 & 0 \end{pmatrix}$$

Figure 2.3 | Example of graph and its adjacency matrix. An undirected weighted graph is represented by a symmetrical adjacency matrix filled with real values which symbolise the magnitude of its edges.

where α is a parameter controlling the rate of diffusion, analogously to the Gaussian kernel. Alternatively, it is possible to define the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is a diagonal matrix whose non-null elements $d(i)$ correspond to the degree of vertex i. A variant is thus the Laplacian Exponential Diffusion Kernel (LEDK):

$$\mathbf{K}_{LED} = \sum_{t=0}^{\infty} \frac{-\alpha^t \mathbf{L}^t}{t!} = e^{-\alpha \mathbf{L}}, \tag{2.5}$$

In these definitions the heat flow is proportional to the number of paths connecting two nodes, so there exist a bias that penalises peripheral nodes with respect to hubs with many interactions. This problem is tackled by a modified version of Eq. 2.5 called Markov Exponential Diffusion Kernel (MEDK) where a $N \times N$ Markov matrix replaces $\mathbf{A}$ [106]:

$$\mathbf{K}_{MED} = e^{\alpha \frac{(N \cdot \mathbf{I} - \mathbf{D} + \mathbf{A})}{N}}. \tag{2.6}$$

Another kernel called Markov Diffusion Kernel (MDK) exploits the model of heat diffusion, but in a different manner [107]. The notion of similarity is based in this case on diffusion patterns, namely on how the heat starting from two nodes v_i and v_j propagates in similar amounts to the other vertices. Within the random walk interpretation, consider the probability of ending up on any other node after a finite number of steps starting from node v_i. A second node v_j is similar to v_i if starting from it a random walker has similar probability to end up on the same nodes after an equal number of steps. Let $\mathbf{P}$ be the transition matrix of a random walk. The MDK is calculated as:

$$\mathbf{K}_{MD} = \mathbf{Z}(t)\mathbf{Z}^T(t) \quad \text{with } \mathbf{Z}(t) = \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{P}^\tau, \tag{2.7}$$

where t is the number of steps of the walk. In order to apply this function to a graph, one only needs to convert its adjacency matrix to a Markov matrix $\mathbf{P}$ by normalising its rows to unitary sum, that is $\sum_j A_{ij} = 1$.

A fourth type of graph node kernel is the Regularized Laplacian Kernel (RLK), which implements a normalised version of the random walk with restart model [108]. The node similarity is defined on the number of paths with different lengths connecting two nodes.

$$\mathbf{K}_{RL} = \sum_{t=0}^{\infty} \alpha^t \left(-\mathbf{L}\right)^t = \left(\mathbf{I} + \alpha\mathbf{L}\right)^{-1}, \tag{2.8}$$

where the exponential discounting rate is α^k instead of $\alpha^k/k!$.

2.3 Multiple kernel learning

Determining the kernel function and its associated parameters that are optimal for a given task can be a major issue. Especially in the case of biological data, it may not be clear what is the best way to estimate the similarity among bio-molecular entities and the decision is usually driven by the available comprehension of the system. Some alternative similarity may better capture the relevant information though. In this situation, a standard procedure is to evaluate the prediction performance of each single kernel and next select that achieving higher performance. However, this evaluation is usually made on a finite number of candidate kernels and the best solution could lie outside the chosen set.

A second big challenge is how to effectively combine kernels when it is necessary to integrate data in heterogeneous formats. As presented in Section 1.4, biological data can assume a variety of structures and formats, which need to be used in concert. Kernel functions align all data types to the Gram matrix formalism, but again there remain various kernels corresponding to different estimates of relatedness.

These challenges can both be tackled by MKL. Given a set of R pre-defined kernels, MKL is a task that aims at finding an optimal kernel combination:

$$\mathbf{K}^{opt} = \psi(\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_R), \tag{2.9}$$

where the Gram matrices $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_R$ provide R feature representations of the data examples. A possible scenario is that in which the kernels correspond to different notions of similarity: instead of selecting a unique best performing similarity, MKL aims at automatically weighting the different similarities and getting a combination of them. Alternatively, the kernels represent various data sources that are processed through appropriate kernel

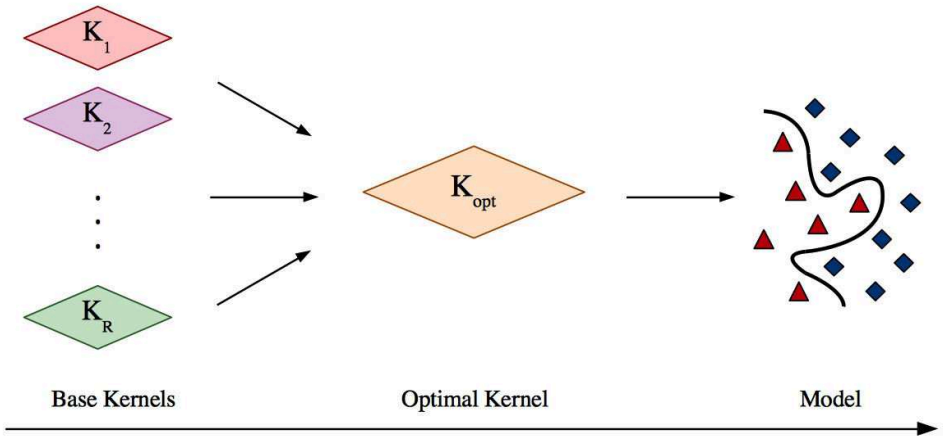


Figure 2.4 | Work-flow of multiple kernel learning. Starting from a set of input kernels, MKL determines an optimal combination which is next used to build a predictive model.

functions. A third scenario sees the first two joined together, namely different data sources have to be combined and for each data source multiple similarities are used (Figure 2.4).

Most often methods focus on learning a positive linear combination of the input kernels, where the function $\psi : \mathbb{R}^R \to \mathbb{R}$ is linear with positive coefficients. In this case the objective is to learn the coefficients $\eta = (\eta_1, \ldots, \eta_R)$ of a conic sum over the Gram matrices:

$$\mathbf{K}^{opt} = \sum_{r=1}^{R} \eta_r \mathbf{K}_r, \ \eta_r \geq 0 . \tag{2.10}$$

The condition of non-negativity on coefficients has an important implication in terms of interpretation: the final feature representation is the rescaled concatenation of the R feature spaces associated to the input kernels. In formula:

$$\phi_{opt}(\mathbf{x}) = \begin{pmatrix} \sqrt{\eta_1}\,\phi_1(\mathbf{x}^1) \\ \sqrt{\eta_2}\,\phi_2(\mathbf{x}^2) \\ \vdots \\ \sqrt{\eta_R}\,\phi_R(\mathbf{x}^R) \end{pmatrix} . \tag{2.11}$$

This allows easily assessing the resulting optimal weights as the relative relevance of each kernel to the task. In particular, in the case of convex sums there is an additional constraint on the weights: $\sum_{r=1}^{R} \eta_r = 1$. Here the weights are bounded between 0 and 1, hence they can be read as percentage of how much each matrix contributes to the final similarity.

2.4 Scalable multiple kernel learning

Recently, many MKL methods have been proposed [90, 91]. However, most of them require a long computation time and a high memory consumption, especially when the number of pre-defined kernels is high. To tackle these limitations, a scalable multiple kernel learning named EasyMKL has been previously proposed [103]. This method focuses on learning a linear combination of the input kernels with positive linear coefficients, like in Equation 2.10. In a fully supervised binary task, EasyMKL computes the optimal kernel by maximising the distance between positive and negative examples. The base learner is an approach for a Kernel-based Optimisation of Margin Distribution (KOMD) in binary classification or ranking [109]. In the following, we will first briefly describe the rationale of KOMD and next present EasyMKL.

Let us first define a set of positive examples $\mathscr{P}$ and a disjoint set of negative examples $\mathscr{N}$. We introduce the probability distribution $\gamma \in \mathbb{R}^N_+$ representing weights assigned to training examples and living in the domain $\Gamma = \{\gamma \in \mathbb{R}^N_+ | \sum_{i\in\mathscr{P}} \gamma_i = 1, \sum_{i\in\mathscr{N}} \gamma_i = 1\}$. From this definition, it follows that any element $\gamma \in \Gamma$ represents a pair of points in the input space: the first one is constrained to the convex hull of positive training examples and the second one to the convex hull of negative training examples. Furthermore, let us define $\mathbf{Y}$ as a diagonal matrix containing the vector of example labels, +1 for the positive and -1 for the negative.

KOMD uses a game-theoretic interpretation of the margin optimisation in a binary task. The problem can be formulated as:

$$\min_{\gamma\in\Gamma} \left\{ (1-\lambda)\gamma^\top \mathbf{YKY}\gamma + \lambda\, \gamma^\top \gamma \right\}. \tag{2.12}$$

Here λ is a parameter controlling the regularisation on the objective function. optimisation of the first term alone leads to an optimal probability distribution γ^{opt} representing the two nearest points in the convex hulls of positive and negative examples, equally to a hard Support Vector Machine (SVM) task using a kernel $\mathbf{K}$. The second term represents a quadratic regularisation over γ whose objective solution is the squared distance between positive and negative centroids in the feature space. The regularisation parameter $\lambda \in [0,1]$ permits to tune the objective to optimise, by balancing between the two critical values $\lambda = 0$ and $\lambda = 1$. When $\lambda = 0$ we obtain a pure hard SVM objective, while when $\lambda = 1$ we get a centroid-based solution.

As stated above, EasyMKL maximises the distance between positive and negative examples, optimising the margin distribution at the same time. Under this notation, the task can be

posed as a min-max problem over variables γ and η as follows:

$$\max_{\eta:\|\eta\|_2\leq 1} \min_{\gamma\in\Gamma} \left\{(1-\lambda)\gamma^\top \mathbf{Y}(\sum_r \eta_r \mathbf{K}_r)\mathbf{Y}\gamma + \lambda\,\gamma^\top\gamma\right\}. \tag{2.13}$$

It can be shown that this problem has analytical solution in the η variable, so that the previous expression can be reshaped into:

$$\min_{\gamma\in\Gamma} \left\{(1-\lambda)\gamma^\top \mathbf{Y}\mathbf{K}^{sum}\mathbf{Y}\gamma + \lambda\,\gamma^\top\gamma\right\}, \tag{2.14}$$

where $\mathbf{K}^{sum} = \sum_r^R \mathbf{K}_r$ is the sum of the pre-defined kernels. This minimisation can be efficiently solved and only requires the sum of the kernels. The computation of the kernel summation can be easily implemented incrementally and only two matrices need to be stored in memory at a time. As shown in [103], EasyMKL can deal with an arbitrary number of kernels using a fixed amount of memory and a linearly increasing computation time.

Once the problem in Eq. 2.14 is solved, we have an optimal distribution γ^{opt} and we are able to obtain the optimal kernel weights η_r^{opt} by using the formula:

$$\eta_r^{opt} = \frac{\gamma^{opt}\mathbf{Y}\mathbf{K}_r\mathbf{Y}\gamma^{opt}}{\sum_{r=1}^R \gamma^{opt}\mathbf{Y}\mathbf{K}_r\mathbf{Y}\gamma^{opt}}. \tag{2.15}$$

The optimal kernel is thus evaluated as $\mathbf{K}^{opt} = \sum_r^R \eta_r^{opt}\mathbf{K}_r$. Finally, by replacing $\mathbf{K}^{sum}$ with $\mathbf{K}^{opt}$ in Eq. 2.14, we can get the final probability distribution γ^{opt}.

2.5 Class-unbalanced multiple kernel learning

In the previous section we introduced EasyMKL, a scalable, efficient kernel integration approach. However, the gene prioritisation task has two additional issues that complicate the work. First, our learning setting is not fully supervised: an assumption is that there are some positive examples hidden among the negatives and we want to retrieve them. Thus, we have the certainty about positive examples but not about negative ones. Second, the number of known disease genes is typically much smaller than the number of candidates, making the problem strongly unbalanced [110]. For these reasons, inspired by a previous work [111] we propose a new MKL algorithm based on EasyMKL that not only inherits its scalability, but also efficiently deals with an unbalanced setting.

Let us consider a set of genes $\mathscr{G} = \{g_1, g_2, \ldots, g_N\}$ that represents either the global set of genes in the genome or a subset of it. Given another set $\mathscr{P} = \{g_1, g_2, \ldots, g_m\}$, $\mathscr{P} \subset \mathscr{G}$ containing genes known to be associated to a genetic disease, gene prioritisation is the

task that aims at ranking genes in the set of candidates $\mathscr{U} = \mathscr{G} \setminus \mathscr{P} = \{g_{m+1}, g_{m+2}, \ldots, g_N\}$ according to their likelihood of being related to that disease. Genes in $\mathscr{P}$ are labelled as *positive* and represent a secure source of information. In contrast, candidate genes in $\mathscr{U}$ are technically *unlabelled*, as we expect that some of them may be associated to the disease but we do not know which ones. Under this notation, the problem can be posed as a Positive-Unlabelled (PU) learning task [95, 97].

In order to clearly present our method, we first need to highlight the different contributions given by positive and unlabelled examples. Therefore, we define $\mathbf{K}^+$, $\mathbf{K}^-$ and $\mathbf{K}^{+-}$ the sub-matrices of $\mathbf{K}^{sum}$ pertaining to positive-positive, unlabelled-unlabelled and positive-unlabelled example pairs, respectively. Schematically, we have:

$$\mathbf{K}^{sum} = \begin{pmatrix} \mathbf{K}^+ & \mathbf{K}^{+-} \\ \mathbf{K}^{+-\top} & \mathbf{K}^- \end{pmatrix},$$

In other words, $\mathbf{K}^+$ contains similarities among positive examples $g_i \in \mathscr{P}, i = 1, \ldots m$, $\mathbf{K}^-$ contains similarities among unlabelled examples $g_j \in \mathscr{U}, j = m+1, \ldots N$ and $\mathbf{K}^{+-}$ includes similarities between positive-unlabelled example pairs. In the same way, we define γ_+ and γ_- as the probability vectors associated to positive and unlabelled examples, respectively.

Under this change of variables, we reformulate the problem as:

$$\min_{\gamma \in \Gamma} \left\{ {\gamma_+}^\top \mathbf{K}^+ \gamma_+ - 2\,{\gamma_+}^\top \mathbf{K}^{+-} \gamma_- + {\gamma_-}^\top \mathbf{K}^- \gamma_- + \lambda_+ {\gamma_+}^\top \gamma_+ + \lambda_- {\gamma_-}^\top \gamma_- \right\}. \tag{2.16}$$

In this new formulation, the original EasyMKL problem is obtained by setting $\lambda_+ = \lambda_- = \frac{\lambda}{1-\lambda}$. However, due to the unbalanced PU nature of the problem, we are interested in using two different regularisations among positive and unlabelled examples. In our case, we decide to fix *a priori* the regularisation parameter $\lambda_- = +\infty$, corresponding to fixing $\lambda = 1$ over unlabelled examples only. Then, the solution of part of the objective function is defined by the uniform distribution $\gamma_- = (\frac{1}{n}, \frac{1}{n}, \ldots \frac{1}{n}) \equiv u$, where $n = N - m$ is the number of unlabelled examples.

We inject this analytic solution of part of the problem in our objective function as follows:

$$\min_{\gamma \in \Gamma^+} \left\{ {\gamma_+}^\top \mathbf{K}^+ \gamma_+ - 2\,{\gamma_+}^\top \mathbf{K}^{+-} u + u^\top \mathbf{K}^- u + \lambda_+ {\gamma_+}^\top \gamma_+ + \lambda_- u^\top u \right\}, \tag{2.17}$$

where $\Gamma^+ = \{\gamma \in \mathscr{R}^m_+ | \sum_{i=1}^m \gamma_i = 1, \gamma_j = 1/n \; \forall \, j = m+1, \ldots N\}$ is the probability distribution domain where the distributions over the unlabelled examples correspond to the uniform distribution. It is trivial that $u^\top \mathbf{K}^- u$ and $\lambda_- u^\top u$ are independent from the γ_+ variable. Then,

they can be removed from the objective function obtaining

$$\min_{\gamma \in \Gamma^+} \left\{ {\gamma_+}^\top \mathbf{K}^+ \gamma_+ - 2\,{\gamma_+}^\top \mathbf{K}^{+-} u + \lambda_+ {\gamma_+}^\top \gamma_+ \right\}. \tag{2.18}$$

In this expression, we only need to consider the entries of the kernel $\mathbf{K}^{sum}$ concerning the positive set, avoiding all the entries with indices in the unlabelled set. The complexity becomes quadratic in the number of positive examples m, which is always much smaller than the number of examples to prioritise. Moreover, this algorithm still depends linearly on the number of kernels R and the overall time complexity is then $\mathcal{O}(m^2 \cdot R)$. In this way, we greatly simplify the optimisation problem, while being able to take into account the diverse amount of noise present in positive and unlabelled example sets. We name the approach here described SCalable UnBAlanced gene prioritization (Scuba).

Like in the previous section, after solving the problem of Eq. 2.18 we use Eq. 2.15 to compute the optimal kernel weights η_r^*. Next, we solve again the Scuba optimisation problem to get the final optimal probability distribution γ^{opt}. Test genes are evaluated by taking the weighted sum over all rectangular test kernel matrices $\mathbf{K}_r^t$, where rows and columns represent test and training genes respectively. In formula:

$$\mathbf{K}^{opt\,t} = \sum_{r=1}^{R} \eta_r^{opt} \mathbf{K}_r^t.$$

The likelihood of association to the disease for any test gene g_i is given by the score s_i defined as

$$s_i = \sum_j y_j \gamma_j^{opt} K_{ij}^{opt\,t}, \tag{2.19}$$

where y_j and γ_j^{opt} are the label and optimal weight of any training example g_j and $K_{ij}^{opt\,t}$ is the optimal kernel value between g_j and the test gene g_i. In other words, s_i is the weighted sum over the similarities between the test gene g_i and all genes in the training set. Once we get the scores for test genes, they can be prioritised based on their score values.

2.5.1 Hierarchical kernel learning

Thanks to the scalability of the new algorithm, it is possible to feed it with numerous base kernels. However, when a high number of parameters are learned, the risk for over-fitting increases. To limit this, one could split the problem into successive optimisations. For instance, consider the scenario where the R kernels represent different kernel functions for various data sources. A strategy could be to learn first the optimal kernel matrix for each

dataset and then learn their optimal combination. In the same way, it is possible to learn the optimal matrix for each kernel function and then learn the optimal combination of the different kernel functions.

Imagine that the set of base kernels can be split into I partitions of J matrices, such that $I \cdot J = R$. Partitions can correspond either to different data sources or kernel functions. During the first phase, we learn R coefficients via I separate tasks using Equation 2.15. This allows getting I intermediate optimal Gram matrices. Next, we feed Scuba with them and obtain the final optimal kernel $\mathbf{K}^{opt}$. In this way, we directly use MKL to perform an automatic selection of optimal kernel parameters. The final kernel and the disease gene set $\mathscr{P}$ are then employed to train a model, which is used to generate a score list for candidate genes in $\mathscr{U}$ through Eq. 2.19. The score assigned to a candidate expresses the likelihood of it being associated to the disease.

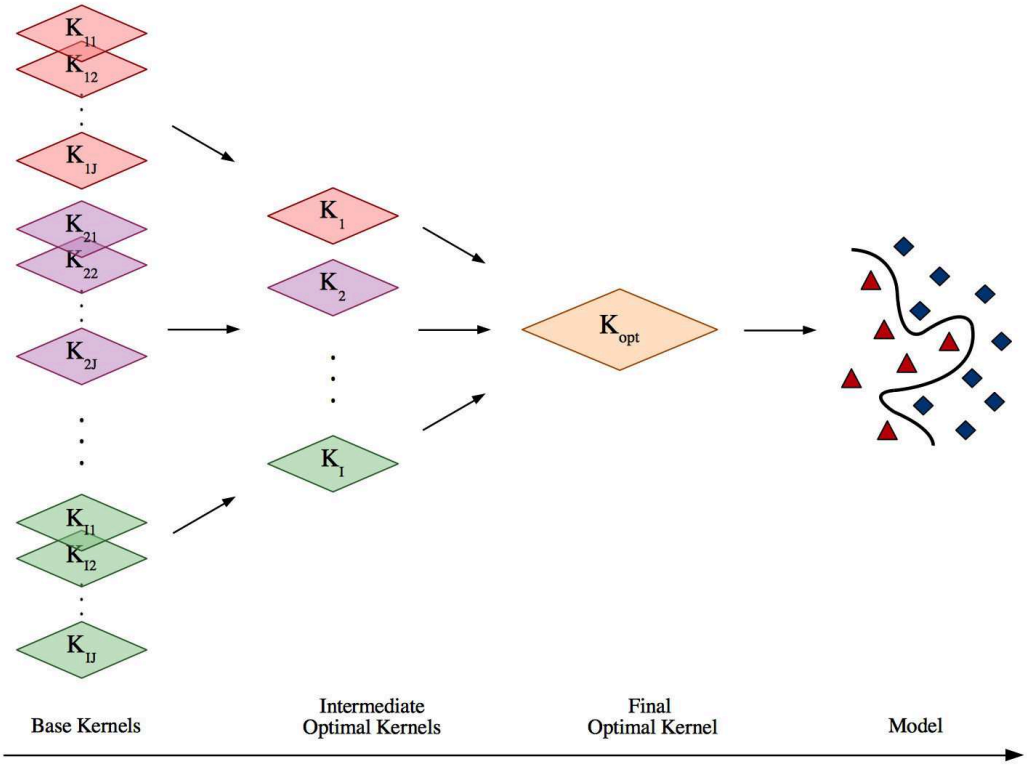


Figure 2.5 | Work-flow of hierarchical multiple kernel learning. Starting from a set of input kernels, hierarchical MKL determines an optimal combination via a two-step procedure. Ideally, input matrices should be grouped according to the data source or the kernel function they represent.

Chapter 3

Predicting disease genes via class-unbalanced multiple kernel learning

The proposed multiple kernel learning method is evaluated by cross-validation and by simulating real case studies. It is observed that, on average, it can effectively prioritise candidate genes ranking on top truly implicated ones. However, prediction improvement is not always achieved when combining heterogeneous datasets and strong biases can exist depending on data sources used.

In this Chapter, we present a series of experimental evaluations on the gene prioritisation method proposed in Chapter 2, called SCalable UnBAlanced gene prioritization (Scuba). In particular, two separate settings are considered. In the first setting, we aim at estimating generalised performance in a standard framework known as cross-validation. In the second setting we evaluate Scuba by means of an approach developed *ad hoc* for gene prioritisation tools, conceived to limit some biases present in cross-validation. Both these assessments are accompanied by comparisons with existing methods and web tools available to researchers, among which we consider two kernel-based methods for gene prioritisation.

In the last Section of the Chapter, we apply Scuba on a case study carried on in collaboration with the BioInfoGen project Unit of the Department of Women's and Children's Health. We prioritise suspect disease genes obtained following Next-Generation Sequencing (NGS) and variant analysis, identifying potential causative genes in the context of renal development disorders.

3.1 Cross-validation

As a first evaluation of Scuba, we followed the experimental protocol used by Chen *et al.* to test predictive performance of gene prioritisation methods [100]. This procedure is based on a popular analysis for the evaluation of predictive algorithms, namely cross-validation [112]. Its rationale is to repeatedly perform training and predictions, masking a different portion of positive genes as unlabelled on each round. The full set of genes $\mathscr{G}$ is randomly split in disjoint subsets of equal size so that at each round a single subset is used for testing and all the remaining for training. The accuracy of an algorithm is estimated on the basis of predictions made on all rounds.

To perform the experiments, we resorted to known gene-disease associations from Online Mendelian Inheritance in Man database (OMIM), grouped into 20 classes on the basis of disease relatedness by Goh *et al* [113]. Among those classes we selected the 12 with at least 30 confirmed genes, excluding cancer (see Table 3.1). We then built a training set consisting of a positive set $\mathscr{P}$ and an unlabelled set $\mathscr{U}$ for each of them. $\mathscr{P}$ contains all its disease gene members. $\mathscr{U}$ is constructed by randomly picking genes from known disease genes such that $|\mathscr{U}| = \frac{1}{2}|\mathscr{P}|$. The unlabelled genes relate to at least one disease class, but do not relate to the considered class. We chose the genes in $\mathscr{U}$ from the other disease genes because we assumed that they were less likely to be associated to the considered class. In fact, disease genes are generally more studied and a potential association has more chances to have already been identified.

After that, a modified version of Leave-One-Out Cross Validation (LOOCV) was used to evaluate Scuba. Iteratively, each gene in the training set was selected to be the test gene and the remaining genes in $\mathscr{P}$ and $\mathscr{U}$ were used to train the model. Once the model was trained, a score list for the test gene and all genes not used in training was computed. Then, we calculated a decision score q_i for each test gene representing the percentage of candidate genes ranked lower than it, using the following formula:

$$q_i = \frac{|\{g_j | s_i \geq s_j\}|}{N}, \tag{3.1}$$

where N in this case is the number of test genes. We collected all decision scores for every gene in any disease class to form disease-specific decision score lists. Scuba accuracy were measured in terms of Area Under the receiver-operating-characteristic Curve (AUC) obtained from any decision score list [114]. The AUC expresses the probability that a randomly chosen disease gene is ranked above a randomly picked non-disease gene for any disease class. An unitary AUC represents a ranking where all truly positive genes are positioned on top, whereas an AUC of 0.5 corresponds to a random predictor.

In this setting, we employed three data sets borrowed from the authors of the benchmark work [100]:

- **HPRD** [115]. The Human Protein Reference Database resource provides protein interaction data which we implement as an unweighted graph, where genes are linked if their corresponding proteins interact.
- **BioGPS** [116]. It contains expression profiles for 79 human tissues, which are measured by an Affymetrix U133A array. Gene co-expression, defined in terms of pairwise expression Pearson Correlation Coefficient (PCC) across all tissues, is used to build an unweighted graph. Any pair of genes is linked by an edge if the PCC value is larger than 0.5, independently of the disease class considered.
- **Pathways**. Pathway datasets are obtained from the database of Kyoto Encyclopedia of Genes and Genomes (KEGG) [73], Reactome [74], PharmGKB [117] and PID [118], which contain 280, 1469, 99 and 2679 pathways, respectively. A pathway co-participation network is constructed by connecting genes that co-participate in any pathway.

These datasets were obtained already preprocessed in such a way that all of them represent exactly the same 7311 genes and were not further processed. We applied the Markov Exponential Diffusion Kernel (MEDK) on them, analogously to the source study, generating three kernel matrices from each network by setting $\alpha = 0.01, 0.04, 0.07$.

In Table 3.1 it is possible to see the performance associated to these three datasets and to their combinations obtained by averaging the corresponding kernels and by Scuba. As it is known, the arithmetic kernel average can be a strong benchmark in some cases, as it is observed here. However, Scuba reaches better AUCs in most trials.

3.2 Time-stamp validation

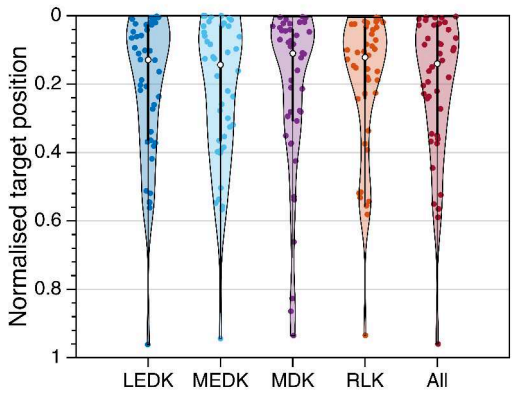
Although the previous evaluation is useful to compare Scuba with other methods, predictive performance in cross-validation experiments may be inflated as compared to real applications. Indeed, the retrieval of known disease genes can be facilitated by various means. One mean is the crosstalk between data repositories: for example, KEGG draws its information also from medical literature [73]. Moreover, often the discovery of the link between a gene and a disease coincides with the discovery of a functional annotation or of a molecular interaction. In practice, instead, researchers are interested in novel associations, which in most cases are harder to find due to a lack of information around them.

Disease class	HPRD	BioGPS	Pathways	Kernels Average	Scuba
Cardiovascular	76.9	64.4	80.3	**81.9**	75.6
Connective	43.7	52.6	74.0	69.2	**78.3**
Dermatological	85.7	86.4	80.1	86.5	**88.6**
Developmental	67.3	54.1	65.3	66.7	**79.7**
Endocrine	71.7	69.5	72.9	**78.6**	78.4
Hematological	79.8	76.6	62.6	73.9	**89.5**
Immunological	89.8	75.8	96.3	**96.4**	**96.4**
Metabolic	79.6	72.8	90.4	90.7	**96.1**
Muscular	66.9	74.0	72.1	75.5	**90.9**
Ophthalmological	70.9	62.0	62.3	72.1	**84.4**
Renal	78.8	76.8	75.7	81.9	**84.0**
Skeletal	75.3	76.8	76.3	**82.8**	77.0
All	76.9	71.6	78.1	81.9	**87.6**

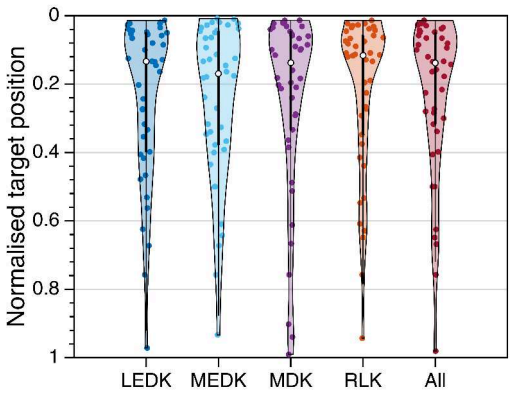
Table 3.1 | AUC for different disease classes. Scuba predictive performance for 12 disease classes in terms of AUC and using different data sources, in the validation setting of Chen *et al* [100]. Bold entries correspond to best performing data sources or their combinations.

In order to achieve a thorough evaluation of Scuba, we tested it in a more realistic setting, following the work of Börnigen *et al* [119]. In this study, eight gene prioritisation web tools were benchmarked as follows. Newly discovered gene-disease associations were collected over a timespan of six months, gathering 42 test genes associated to a range of disorders. As soon as a new association was discovered, each web tool was queried with a disorder-specific set of positive genes $\mathscr{P}$ to prioritise a set of candidates $\mathscr{U}$ containing the corresponding test gene (or to prioritise the whole genome where possible). Rank positions of the 42 test genes were ultimately used to assess the ability of the tools to successfully prioritise disease genes. The idea behind this procedure is to anticipate the integration of the associations in the data sources and so avoid biased predictions.

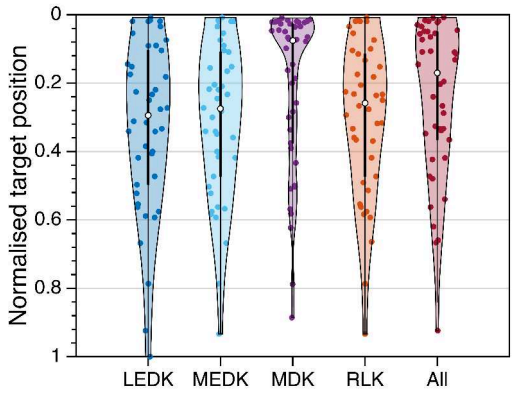
In order to test Scuba in this setting, we backdated our data to a time prior to May 15, 2010 by employing String v8.2 data [80]. We generated three kernel matrices for each of the four functions described in Section 2.2.2, setting the parameters as follows: $\alpha = 0.01, 0.04, 0.07$ for Laplacian Exponential Diffusion Kernel (LEDK) and Markov Exponential Diffusion Kernel (MEDK), $t = 2, 4, 6$ for Markov Diffusion Kernel (MDK) and $\alpha = 1, 10, 100$ for Regularized Laplacian Kernel (RLK). After that, we recovered positive sets and test genes from the original publication and we followed its experimental protocol as follows [119]. We performed prioritisations for each test gene in two distinct cases: genome-wide and candidate set-based prioritisations. In any genome-wide prioritisation all genes in the String dataset - except those in $\mathscr{P}$ - belong to $\mathscr{U}$ and were prioritised. In any candidate set-based



(a)



(b)



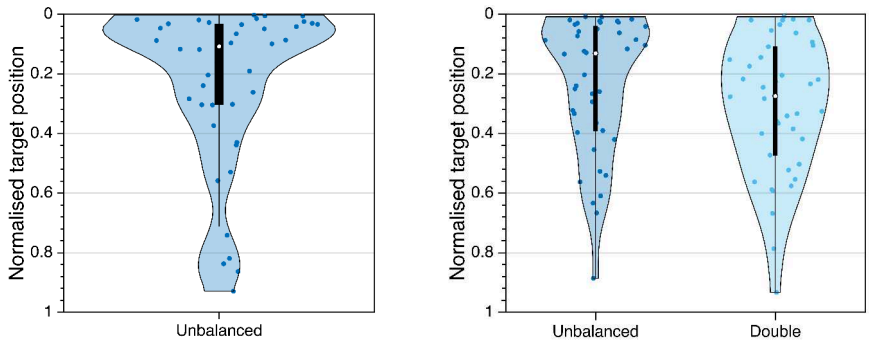
(c)

Figure 3.1 | Test genes rank distributions in simulated case studies - (1/2). Comparison of rank distributions predicted by Scuba for the 42 test genes as a function of the kernel applied on the String network (see Section 2.2.2): **(a)** genome-wide prioritisations; **(b)** candidate set-limited prioritisations; **(c)** candidate set-limited prioritisations when optimising both regularisation hyper-parameters (Eq. 2.16).

prioritisation, the set of candidates $\mathscr{U}$ was constructed by considering all genes with Ensembl gene identifier within the chromosomal regions around the test gene, in order to get on average 100 candidates [120].

Figures 3.1a, 3.1b show the rank distribution for the 42 test genes in these two scenarios, respectively, for the different kernel functions and their simultaneous combination. Figure 3.1c shows instead the same distribution for candidate set-based prioritisations when resorting to Scuba with two free regularisation hyper-parameters λ_+ and λ_-, corresponding to the optimisation problem of Eq. 2.16. In this last case, training was performed using only disease and candidate genes, in order to have a viable number of parameters to learn. Overall, MDK appears the best performing kernel.

Furthermore, we tested the hierarchical Multiple Kernel Learning (MKL) strategy introduced in Section 2.5.1. In this case, we limit ourselves to the case where we combine all kernel functions together. Figure 3.2a shows that the genome-wide test gene rank distribution



(a) **(b)**

Figure 3.2 | Test genes rank distributions in simulated case studies - (2/2). Comparison of rank distributions predicted by the hierarchical version of Scuba for the 42 test genes: **(a)** genome-wide prioritisations; **(b)** candidate set-limited prioritisations adopting an unbalanced regularisation with $\lambda_- = +\infty$ or adopting a double free regularisation.

is more pushed toward the extremes as compared to classical MKL employed in Figure 3.1a: the median is higher but the tail of the distribution is also more populated.

In this setting we also investigated a bias inherent to the String dataset. As shown in previous studies, network-driven functional predictions are influenced by node degree, with a tendency by highly connected genes to obscure more peripheral ones [121]. In practise, this means that strongly linked genes tend to be ranked higher than relatively isolated ones. Generally such hubs are deeply studied genes or multi-functional genes involved in many biological processes, but at the same time not what a researcher may be looking for. To verify the presence of this bias in our prioritisations, we measured to what extent the rank of test genes is related to the number and the strength of their connections in the String network. In other words, for each kernel function we calculated the SCC ρ between the genome-wide rank of target genes and the sum of their edge weights. We obtained a strikingly high correlation for all kernel functions, as it visible in Figure 3.3. The only function that appears to contrast this behaviour to any degree is MDK, with a SCC significant to a 5% threshold but not at a 1% threshold. This is consistent with its higher performances.

3.3 Comparison with competing methods

In this Section, we show how Scuba performs as compared to existing methods for gene prioritisation. Table 3.2 illustrates the performance of different techniques in this experimental setting reported by Chen *et al* presented in the previous Section [100]. In the second column

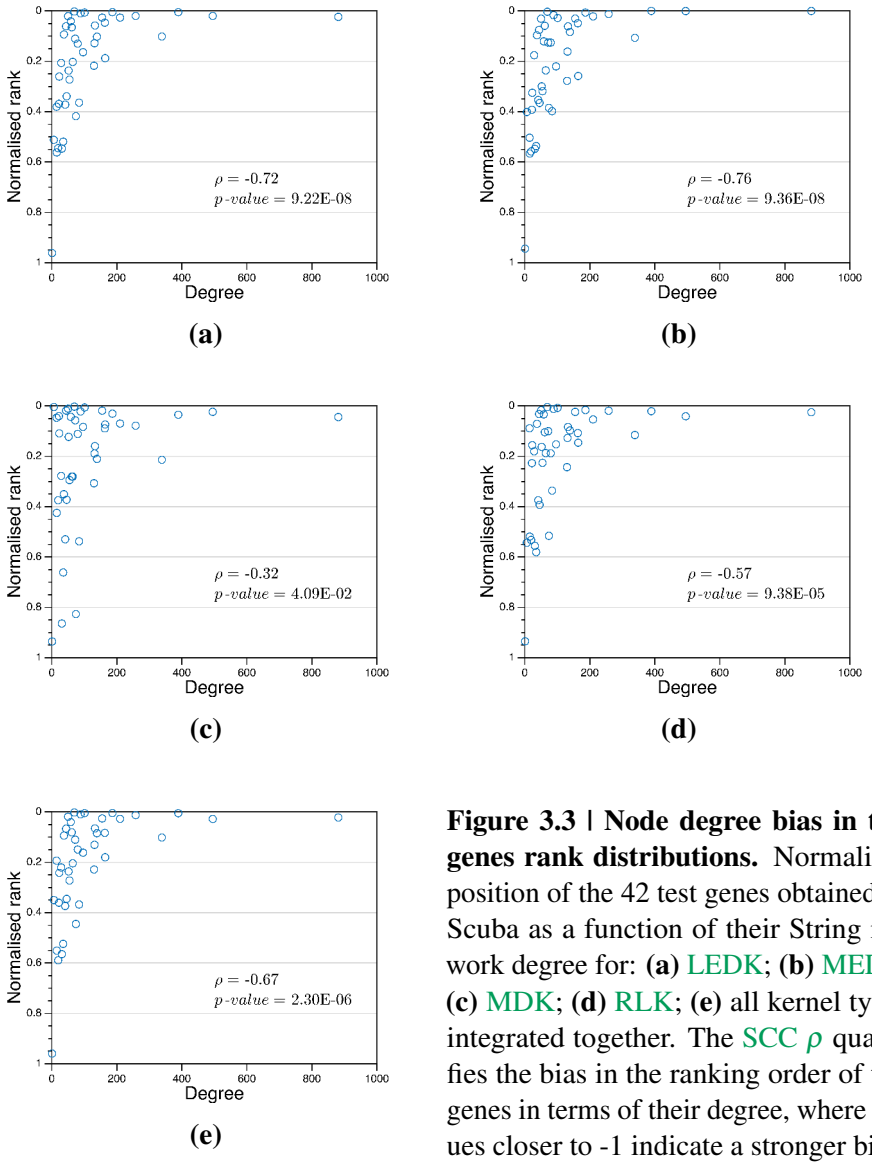


Figure 3.3 | Node degree bias in test genes rank distributions. Normalised position of the 42 test genes obtained by Scuba as a function of their String network degree for: **(a)** LEDK; **(b)** MEDK; **(c)** MDK; **(d)** RLK; **(e)** all kernel types integrated together. The SCC ρ quantifies the bias in the ranking order of test genes in terms of their degree, where values closer to -1 indicate a stronger bias.

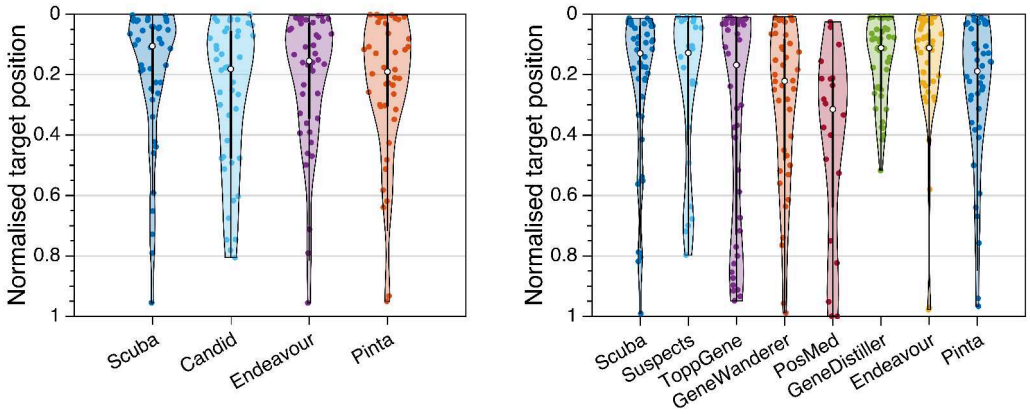
we show the significance of the difference between reported AUCs and Scuba AUC, assessed by means of separate statistical tests to control the comparison-wise error rate [114, 122]. It can be seen that Scuba performs significantly better than the other methods.

As regards the time-stamp validation, in Figure 3.4 and Table 3.3 we show results for Scuba compared to the results obtained in the work of Börnigen *et al*, pertaining to eight prioritisation systems [119]. In this case we resorted to a combination of MDK and RLK matrices with the same parameter values as before. To realise a meaningful comparison, we calculated median, mean and standard deviation of the normalised ranks for test genes. We also computed the True Positive Rate (TPR) relatively to some representative

Method	AUC	p-value
Scuba	**0.876**	-
F3PC [100]	0.830	$1.39 \cdot 10^{-4}$ *
MRF [106]	0.731	$<10^{-6}$ *
DIR [123]	0.716	$<10^{-6}$ *
GeneWanderer [124]	0.711	$<10^{-6}$ *

Table 3.2 | AUC comparison. Performance of different techniques in the experimental setting of Chen *et al* [100] expressed in terms of AUC. Except for our proposed method Scuba, these results were taken from that work. The p-values indicate significance of the pairwise AUC differences compared to Scuba AUC [114]. Asterisks indicate significance of the test (p-value < 0.05).

thresholds (5%, 10% and 30% of the ranking) and the AUC obtained by averaging over the 42 prioritisations. In genome-wide predictions, Scuba dominates over the other tools. On predictions over smaller candidate sets, it is still competitive although best results are achieved by GeneDistiller [125], Endeavour [88] and ToppGene [126]. It is important to underline that in this case considered tools rely on different data sources, so we are comparing different prioritisation systems rather different algorithms. Furthermore, tools are in some cases unable to provide an answer to a given task, depending on the underlying data sources (for more details see the original work [119]). We report the fraction of prioritisations on which tools are actually evaluated as response rate. This analysis has the purpose of showing



(a) **(b)**

Figure 3.4 | Rank distribution comparison between Scuba and web tools in simulated case studies. Distributions of the 42 test genes for Scuba and main prioritisation web tools in: **(a)** genome-wide prioritisations **(b)** candidate sets-limited prioritisations.

the potentiality of Scuba relatively to what is easily accessible by non-bioinformaticians. However, since we used the String data source Scuba is directly comparable with Pinta [119, 127].

Along with Scuba, we evaluated in this setting also MKL1class and ProDiGe, two state-of-the-art kernel based gene prioritisation methods described in Section 1.5.1 [94, 95]. We ran ProDiGe using the default parameters indicated in the corresponding paper: number of bagging iterations $B = 30$ and regularisation parameter $C = 1$. In the same way, we set the regularisation parameter $\nu = 0.5$ for MKL1class. In Table 3.4 it is possible to see performances for all three methods. The significance of rank median differences between Scuba and competing methods was assessed by separate Wilcoxon signed rank tests where we control the comparison-wise error rate [122]. At a significance threshold of 5%, Scuba achieves significantly higher performances in genome-wide tasks compared to both baselines. In the candidate set-based setting, it performs significantly better than ProDiGe and better, although not significantly, than MKL1class. These differences can be visually appreciated in Figure 3.5, where we compare the rank distributions obtained by the three methods. Scuba and MKL1class present moderate rank differences, particularly in the central region of the ranks. On the other hand, differences between Scuba and ProDiGe are smaller (Pearson $r = 0.98$ in both cases) and almost all in favour of Scuba.

Next, we expanded this validation by employing gene-phenotype annotations derived from the Human Phenotype Ontology (HPO) [71]. This resource gathers information from several databases and makes available its monthly updates, permitting to trace the annotations history. We downloaded the HPO build 29 - dating March 2013 - and build 117 of February 2017. We compared the two annotations corresponding to these versions of HPO and extracted the gene-phenotype associations that were added in this time gap. We concentrated on abnormal phenotypes associated to the multi-factorial diseases covered in the previous analysis, which could possibly have some previously undiscovered associations. We thus analysed how the obtained genes are ranked in genome-wide prioritisations of the previous analysis, applying the same performance measures as before. The outcome is an analogous evaluation, but this time target genes are those extracted from HPO.

In Table 3.5 results for Scuba, MKL1class and ProDiGe are shown. We can observe a slightly different trend compared to previous results, with Scuba and ProDiGe having very close performance and MKL1class being significantly worse than Scuba. As a confirmation, in Figure 3.5 it is possible to see that there is no clear difference between Scuba and ProDiGe rank distributions. Instead, MKL1class ranks several test genes neatly lower compared to Scuba, with the associated Pearson correlation coefficient dropping to $r = 0.85$.

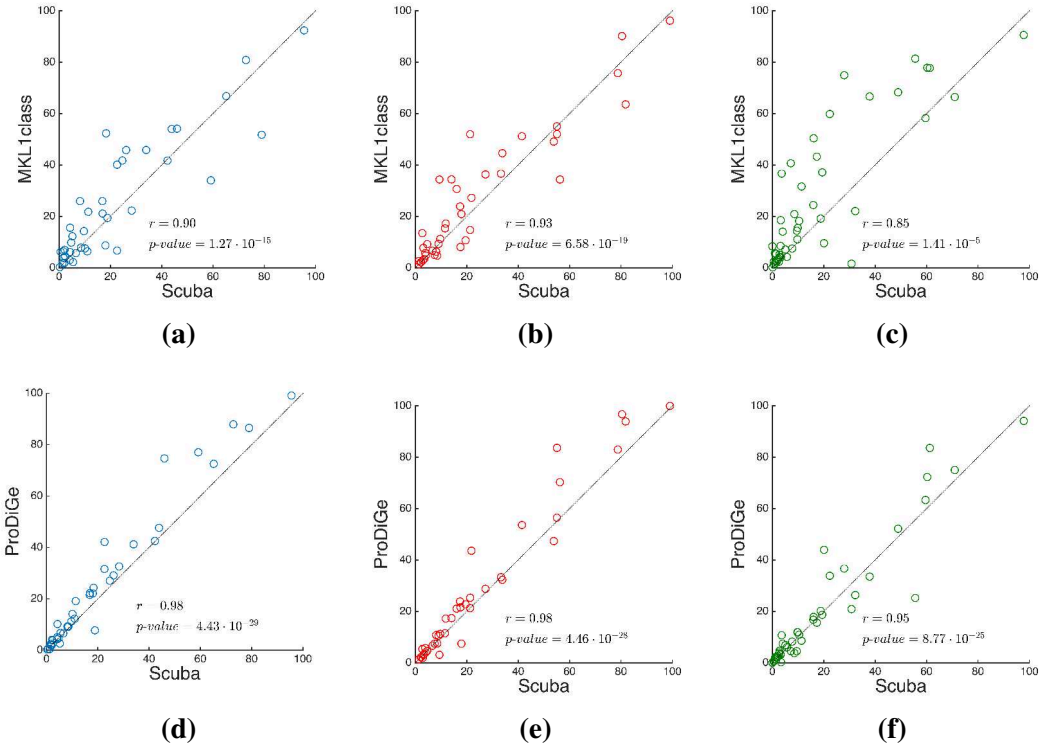


Figure 3.5 | Comparison of rank distributions between Scuba and competing kernel methods. Normalised rank distributions predicted by Scuba, MKL1class and ProDiGe for test genes in **(a)** genome-wide prioritisations in the time-stamp validation of Table 3.4 - **(b)** candidate set-based prioritisations in the time-stamp validation of Table 3.4 - **(c)** genome-wide prioritisations in the expanded time-stamp validation of Table 3.5. In all cases, each point represents a test gene and lower values on the axes indicate better predictions. Genes lying on a diagonal have the same rank according to both methods considered on a plot. The further a gene lies above (below) a diagonal and the better it was ranked by Scuba (MKL1class/ProDiGe) compared to MKL1class/ProDiGe (Scuba). In each plot we show the PCC r between the rank distributions and its associated p-value.

Tool/Method	Response rate (%)	Rank median	Rank average	TPR in top 5% (%)	TPR in top 10% (%)	TPR in top 30% (%)	AUC
Genome-wide prioritisation methods							
Scuba	100	**10.55**	**20.48**±23.53	**33.3**	**47.6**	**78.6**	**0.80**
Candid [128]	100	18.10	27.35±24.62	21.4	33.3	64.3	0.73
Endeavour [88]	100	15.49	21.47±22.37	28.6	38.1	71.4	0.79
Pinta [127]	100	19.03	23.52±23.58	26.2	31.0	71.4	0.77
Candidate set-based prioritisation methods							
Scuba	100	12.95	23.32±25.46	28.6	45.2	73.8	0.78
Suspects [129]	88.9[a]	12.77[a]	24.64±26.42[a]	33.3[a]	33.3[a]	63.0[a]	0.76[a]
ToppGene [126]	97.6	16.80	34.53±35.31	**35.7**	42.9	52.4	0.66
GeneWanderer-RW [124]	88.1	22.10	29.55±26.28	16.7	26.2	61.9	0.71
Posmed-KS [130]	47.6	31.44	42.07±30.98	4.7	7.1	23.8	0.58
GeneDistiller [125]	97.6	**11.11**	**15.37**±13.77	26.2	**47.6**	78.6	**0.85**
Endeavour [88]	100	11.16	18.41±21.39	26.2	42.9	**90.5**	0.82
Pinta [127]	100	18.87	25.23±24.72	28.6	31.0	71.4	0.75

Table 3.3 | Comparison of rank distribution statistics between Scuba and web tools. Performances of Scuba and of main gene prioritisation web tools in the time-stamp validation setting of Börnigen *et al* [119]. Response rate is the percentage of gene-disease associations considered by each tool. Values for Suspects were computed on the first 27 associations only (highlighted by [a]).

Tool/Method	Rank median	Rank average	TPR in top 5% (%)	TPR in top 10% (%)	TPR in top 30% (%)	AUC	Rank difference p-value
Genome-wide prioritisation methods							
Scuba	**10.55**	**20.48**±23.53	**33.3**	**47.6**	**78.6**	**0.80**	-
MKL1class [94]	13.30	23.42±23.23	21.4	**47.6**	69.0	0.77	$2.5 \cdot 10^{-2}$ *
ProDiGe [95]	11.73	24.45±27.33	31.0	45.2	71.4	0.76	$3.0 \cdot 10^{-7}$ *
Candidate set-based prioritisation methods							
Scuba	**12.95**	**23.32**±25.46	**28.6**	**45.2**	**73.8**	**0.78**	-
MKL1class [94]	15.07	25.63±24.73	23.8	40.5	61.9	0.76	$9.7 \cdot 10^{-2}$
ProDiGe [95]	14.41	26.39±29.09	26.2	40.5	71.4	0.75	$2.7 \cdot 10^{-3}$ *

Table 3.4 | Comparison of rank distribution statistics between Scuba and competing kernel methods - (1/2). Performances of Scuba, MKL1class and ProDiGe in the time-stamp validation setting of Börnigen *et al* [119]. Values refer to predictions on all the 42 gene-disease associations. Rank difference p-values were obtained using Wilcoxon signed rank tests comparing separately Scuba/MKL1class and Scuba/ProDiGe ranks differences. Asterisks indicate significance of the tests at a threshold of 5%.

Method	Rank median	Rank average	TPR in top 1% (%)	TPR in top 5% (%)	TPR in top 10% (%)	TPR in top 30% (%)	AUC	Rank difference p-value
Genome-wide prioritisations								
Scuba	8.13	**17.45**±22.33	**10.4**	41.7	**58.3**	**79.2**	**0.83**	-
MKL1class [94]	14.28	25.79±26.96	2.1	27.1	45.8	66.7	0.74	$1.2 \cdot 10^{-5}$ *
ProDiGe [95]	**7.89**	18.40±23.77	**10.4**	**43.8**	54.2	**79.2**	0.82	$9.5 \cdot 10^{-2}$

Table 3.5 | Comparison of rank distribution statistics between Scuba and competing kernel methods - (2/2). Performances of Scuba, MKL1class and ProDiGe in the expanded time-stamp validation setting involving seven multi-factorial diseases. Values refer to predictions on 48 gene-disease associations. Rank difference p-values were obtained using Wilcoxon signed rank tests comparing separately Scuba/MKL1class and Scuba/ProDiGe ranks differences. Asterisks indicate significance of the tests at a threshold of 5%.

3.4 Case study: prioritisation of candidate renal hypo/dysplasia genes

Renal Hypo/Dysplasia (RHD) is a defect in the number and differentiation of nephronic units with a subsequent impairment of kidney function. The most obvious phenotypic trait is a reduction of kidney size often associated with abnormally developed structures (dysplasia). The principal cause is thought to be a perturbation of the nephrourogenetic program, a complex process regulated by a space-time-corrected sequential activation of a cascade of genes. This congenital anomaly, isolated or associated with urinary tract anomalies, account for 20-30% of all causes of chronic renal failure in children, that need renal replacement therapy and kidney transplantation.

Mutations on HNF1B, PAX2, SALL1, SIX1, SIX2, BMP4, EYA1, UMOD and RET genes were identified in non-syndromic RHD patients but justify only a small proportion of sporadic or familial cases [131]. This is deemed due to a larger heterogeneity of locus and to a greater complexity of the pathogenetic mechanisms involved in the determination of RHD. On the other hand, single-gene mutations may cause a wide phenotypic spectrum of features that ranges from vesicoureteral reflux to renal agenesis. In fact, the induction of mesenchyme morphogenesis by the ureteric bud influences morphogenesis of all three tissue groups of the excretory system: the ureterovesical junction, the ureter, and the kidney [132]. Literature studies report mutations in a single gene related to variable phenotypes even within the same family, suggesting the existence of modifier genes. Furthermore, the identification of predisposing genetic and environmental factors in RHD cases points to the possibility of a multi-factorial aetiology for this developmental abnormality. RHD belongs to the Congenital Abnormalities of Kidney and Urinary Tract (CAKUT) disorder class, occurring in 3-6 per 1000 live births and accounting for approximately 30% of all

Gene name	Ensembl identifier
BICC1	ENSG00000122870
BMP4	ENSG00000125378
BMP7	ENSG00000101144
CHD1L	ENSG00000131778
CHD7	ENSG00000171316
CHRM3	ENSG00000133019
DSTYK	ENSG00000133059
FGFR1	ENSG00000077782
FRAS1	ENSG00000138759
FREM1	ENSG00000164946
FREM2	ENSG00000150893
GATA3	ENSG00000107485
GDNF	ENSG00000168621
GFRA1	ENSG00000151892
GLI3	ENSG00000106571
GPC3	ENSG00000147257
HNF1B	ENSG00000275410
ITGA3	ENSG00000005884
ITGA8	ENSG00000077943
JAG1	ENSG00000101384
JAG2	ENSG00000184916
KIP2	ENSG00000136425
MNX1	ENSG00000130675
NIPBL	ENSG00000164190
NOTCH1	ENSG00000148400
NOTCH2	ENSG00000134250
PAX2	ENSG00000075891
PEX1	ENSG00000127980
RET	ENSG00000165731
ROBO2	ENSG00000185008
SALL1	ENSG00000103449
SALL4	ENSG00000101115
SIX1	ENSG00000126778
SIX2	ENSG00000170577
SIX5	ENSG00000177045
SLIT2	ENSG00000145147
TBX3	ENSG00000135111
UPK3A	ENSG00000100373
WNT4	ENSG00000162552
WT1	ENSG00000184937

Table 3.6 | Full list of seed RHD genes used in the prioritisation. A total of 40 genes associated to RHD were employed as positive set.

developmental anomalies. Overall CAKUT show different phenotypes, with a wide spectrum of malformations ranging from unilateral vesico-ureteral reflux to bilateral renal agenesis.

The BioInfoGen Unit of Women's and Children's Health is involved in the study of RHD with the goal of elucidating its genetic bases. NGS of exome coding regions, intron-exon junctions and part of the untranslated regions for twenty undiagnosed patients was performed and lead to the identification of missense or nonsense variants in 2030 genes. In order to guide the following experimental research, we prioritised genes harbouring those mutations via Scuba, resorting to three data sources:

- **String** This source is the version 10.1 of the String database used in the previous section.
- **BioPlex** As briefly described in Section 1.4, the BioPlex 2.0 network is the result of a recent large-scale affinity purification-mass spectrometry investigation embracing previously unmapped portions of the human interactome [56].
- **CPDB** Consensus Path DataBase (CPDB) unifies pathway annotations from all main repositories within a comprehensive database including 3892 biological pathways.

We transformed String and Bioplex by the MDK as before, setting $t = 2, 4, 6$. CPDB was converted into three Gaussian kernel matrices using $\sigma = 0.5, 1, 2$. We first performed prioritisations employing each data source separately, recording the average 10 fold cross-validation AUC in the parameter selection phase. We then integrated the three datasets to achieve the final ranking. As can be seen in Table 3.7, integration of the three datasets improves validation performance on known RHD genes.

Top thirty genes in the final ranking are supported by a much higher score as compared to the other candidates (Figure 3.6). In particular, all top twenty genes are strictly related to embryonic development: they are homeobox genes, transcriptional factors, cellular adhesion growth factors and genes regulating the Wnt signaling pathway, which is important in signal

String	BioPlex	CPDB	AUC
✓			0.87
	✓		0.75
		✓	0.66
✓	✓	✓	0.89

Table 3.7 | Validation on known renal hypo/dysplasia genes. Maximal average AUC obtained by 10 fold cross-validation on known RHD genes during the selection of λ_+.

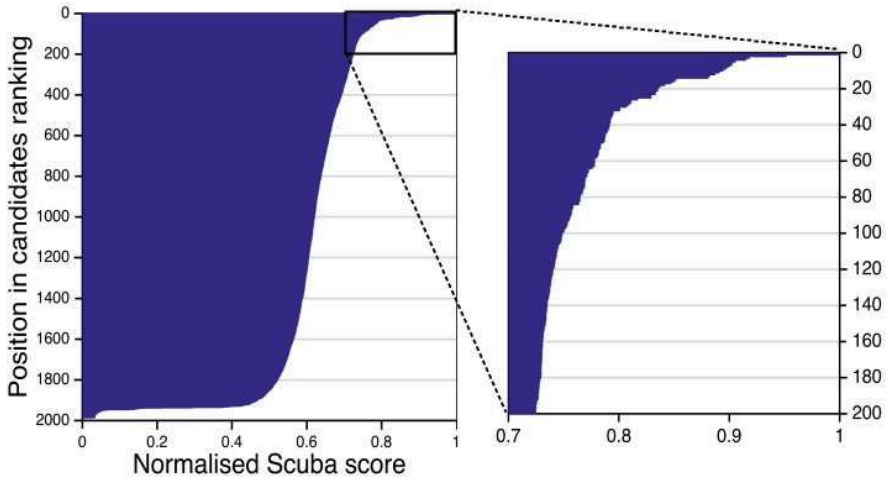


Figure 3.6 | Score distribution for candidate RHD genes. Graphical visualisation of the Scuba score distribution for candidate genes obtained through Eq. 2.19 and normalised in such a way to be constrained between 0 and 1. Most promising candidates have a normalised score close to 1.

transduction for cell faith control, cellular proliferation and migration. Homeobox genes encode DNA-binding proteins, many of which are thought to be involved in early embryonic development. Two genes, FGF20 and TNXB, are known among causative genes for renal agenesis and vesicoureteral reflux [133, 134]. Moreover, 9 genes belong to families known to be involved in CAKUTs. In particular, EYA2 is a EYA1 homologous, which is associated to a syndromic form of CAKUT with a renal phenotype of RHD. Additionally, it seems to be implicated in the regulation of important transcription factors such as SIX2, also involved in the RHD determination also isolated, like that of analysed patients [135]. Overall, this preliminary screening on the top candidates seems to indicate that the achieved prioritisation is promising to identify novel causative genes, as future investigations will verify.

Table 3.8 | Top twenty candidate genes for renal hypo/dysplasia.

Position in candidates ranking	Ensembl identifier	Gene name	Position in global ranking	Description
1	ENSG00000162490	DRAXIN	4	Protein coding gene involved in dorsal inhibitory axon guidance and β-catenin-dependent Wnt Signalling.
2	ENSG00000064655	EYA2	18	Functions both as protein phosphatase and as transcriptional co-activator for SIX1 and probably also for SIX2, SIX4 and SIX5.
3	ENSG00000258873	DUXA	54	Member of the DUXA homeobox gene family. Multiple, related processed pseudo-genes have been found, thought to reflect expression of this gene in the germ line or in embryonic cells.
4	ENSG00000144355	DLX1	65	Member of a homeobox transcription factor gene family, similiar to the Drosophila distal-less gene. The encoded protein is localised to the nucleus where it may function as a transcriptional regulator of signals from multiple TGF-β super-family members, as well as play a role in the control of craniofacial patterning and the differentiation and survival of inhibitory neurons in the forebrain. Alternatively spliced transcript variants encoding different isoforms have been described.
5	ENSG00000135925	WNT10A	90	The WNT gene family consists of structurally related genes which encode secreted signalling proteins. These proteins have been implicated in oncogenesis and in several developmental processes, including regulation of cell fate and patterning during embryogenesis.
6	ENSG00000089225	TBX5	100	Member of a phylogenetically conserved family of genes that share the T-box DNA-binding domain and encode transcription factors involved in the regulation of developmental processes. This gene is closely linked to related family member T-box 3 (ulnar mammary syndrome) on human chromosome 12.
7	ENSG00000152785	BMP3	102	This gene encodes a secreted ligand of the TGF-β (transforming growth factor-β) super-family of proteins. Ligands of this family bind various TGF-β receptors leading to recruitment and activation of SMAD family transcription factors that regulate gene expression. The pre-proprotein is proteolytically processed to generate each subunit of the disulfide-linked homodimer.
8	ENSG00000197757	HOXC6	108	Member of the homeobox gene family, encoding a highly conserved family of transcription factors that play an important role in morphogenesis in all multicellular organisms.
9	ENSG00000104371	DKK4	116	This gene encodes a protein that is a member of the Dickkopf family, contains two cysteine rich regions and is involved in embryonic development through its interactions with the Wnt signalling pathway. Activity of this protein is modulated by binding to the Wnt co-receptor and the co-factor Kremen 2.
10	ENSG00000138675	FGF5	125	The protein encoded by this gene is a member of the fibroblast growth factor (FGF) family. FGF family members possess broad mitogenic and cell survival activities and are involved in a variety of biological processes including embryonic development, cell growth, morphogenesis, tissue repair, tumour growth and invasion.
11	ENSG00000078579	FGF20	133	Member of the FGF family. The gene product is a secreted neurotrophic factor but lacks a typical signal peptide. It is expressed in normal brain, particularly the cerebellum, and may regulate central nervous system development and function. Homodimerisation of this protein was shown to regulate its receptor binding activity and concentration gradient in the extracellular matrix.
12	ENSG00000118257	NRP2	144	This gene encodes a member of the neuropilin family of receptor proteins, a transmembrane protein that binds to SEMA3C protein and SEMA3F protein and interacts with vascular endothelial growth factor. It may play a role in cardiovascular development, axon guidance, and tumorigenesis.

13	ENSG00000107984	DKK1	164	This gene encodes a protein that is a member of the Dickkopf family, contains two cysteine rich regions and is involved in embryonic development through its interactions with the Wnt signalling pathway.
14	ENSG00000139269	INHBE	166	This gene encodes a secreted ligand of the TGF-β (transforming growth factor-β) super-family of proteins. The pre-proprotein is proteolytically processed to generate an inhibin beta subunit, implicated in regulating numerous cellular processes including cell proliferation, apoptosis, immune response and hormone secretion.
15	ENSG00000070018	LRP6	257	This gene encodes a member of the low density lipoprotein (LDL) receptor gene family, transmembrane cell surface proteins involved in receptor-mediated endocytosis of lipoprotein and protein ligands. The protein encoded by this gene functions as a receptor or, with Frizzled, as a co-receptor for Wnt and thereby transmits the canonical Wnt/beta-catenin signaling cascade. Through its interaction with the Wnt/beta-catenin signalling cascade this gene plays a role in the regulation of cell differentiation, proliferation, and migration and the development of many cancer types.
16	ENSG00000168477	TNXB	262	This gene encodes a member of the tenascin family of extracellular matrix glycoproteins, which have anti-adhesive effects. It localises to the major histocompatibility complex (MHC) class III region on chromosome 6 and it is one of four genes in this cluster which have been duplicated. The encoded protein is thought to function in matrix maturation during wound healing and its deficiency has been associated with the connective tissue disorder Ehlers-Danlos syndrome.
17	ENSG00000256463	SALL3	270	This gene encodes a sal-like C2H2-type zinc-finger protein and belongs to a family of evolutionarily conserved genes found in species as diverse as Drosophila, C. Elegans, and vertebrates. Mutations in some of these genes are associated with congenital disorders in human, suggesting their importance in embryonic development. The encoded protein binds to DNA methyltransferase-3-α (DNMT3A) and reduces DNMT3A-mediated CpG island methylation.
18	ENSG00000221818	EBF2	288	The protein encoded by this gene belongs to the COE (Collier/Olf/EBF) family of non-basic, helix-loop-helix transcription factors that have a well conserved DNA binding domain. The COE family proteins play an important role in variety of developmental processes.
19	ENSG00000100060	MFNG	304	This gene is a member of the fringe gene family which also includes radical and lunatic fringe genes. They all encode evolutionarily conserved secreted proteins that act in the Notch receptor pathway to demarcate boundaries during embryonic development.
20	ENSG00000119699	TGFB3	308	This gene encodes a secreted ligand of the TGF-β superfamily of proteins. Ligands of this family bind various TGF-β receptors leading to recruitment and activation of SMAD family transcription factors that regulate gene expression. This protein is involved in embryogenesis and cell differentiation, and may play a role in wound healing.

Chapter 4

Data integration via constraint-based modelling

Constraint-based modelling offers a flexible framework for the analysis of genotype-phenotype interactions on a systems scale and with a single reaction resolution. In particular, it allows the full reconstruction of the metabolic states associated to genetic perturbations. Data-driven gene prioritisation can thus be implemented upon patterns of metabolic rearrangements, which unify alterations at multiple omics levels.

Metabolism is one of the major biological components that co-participates with the genotype in composing the phenotype. As mentioned in Section 1.1, metabolic adjustments can compensate or modify genetic alterations, through complex non-intuitive routes. Incorporating metabolic information in the prediction of disease genes could therefore plausibly unveil novel insights. However, in comparison to other popular omics, large-scale metabolic data acquisition is still immature and suffers from major limitations. The main obstacles are a high biochemical heterogeneity and concentration variations that can occur within sub-second time scales and span several orders of magnitude [136, 137].

Genome-Scale Metabolic Models (GSMMs) are mathematical representations of all known biochemical reactions and their associated enzymes and encoding genes that comprise the metabolic functionality of a cell [138]. A vast range of computational methods have been developed upon GSMMs to investigate interactions between genotype, environment and phenotype [23, 139]. Acting as integrative platforms for multi-omics data, they can also help identify non-intuitive phenomena in metabolism [140]. Importantly, they also permit to evaluate the complete metabolic state of cell populations even when metabolome profiling is infeasible.

The mathematical framework of GSMMs is Constraint-Based Modelling (CBM) and lies on two main assumptions. First, mass and charge conservation, as stated by classical physics laws. This guarantees that the total mass of produced substrates equals the total mass of consumed substrates. Second, the system must be at steady state, meaning that reaction rates do not change over time. The steady state assumption differentiates CBM from the modelling based on ordinary differential equations. The latter has the advantages to be very precise and to allow studying metabolic systems in dynamical conditions. On the other hand, it is very computationally expensive and it also requires to know in detail initial metabolic conditions and kinetic reaction parameters. For these reasons, it is feasible only for small systems so it cannot capture long range phenomena or general metabolic reprogramming. Conversely, GSMMs are restricted to steady-state conditions, but can span the whole cellular metabolism or even multi-cellular communities [141].

CBM development has been fuelled by whole-genome sequencing, as it lies on the knowledge of genotype and on the functional or biochemical annotation of its products. The first GSMM to be assembled was of Haemophilus influenzae Rd, which was also the first organism having an established genomic sequence [142]. Soon after, the achievement of the Escherichia Coli genome permitted the construction of its metabolic map and models for several other bacterial organism were generated ever since [143]. The first global reconstruction of human metabolism - Recon 1 - dates 2007 [144], while the most improved versions to date are Recon 2.2 and the Human Metabolic Reconstruction 2 [29, 30]. Insights into human disorders can be gained by means of these GSMMs, both by testing specific hypothesis at a single reaction resolution and by analysing global cellular reprogramming [145–147]. In particular, the combined application of CBM and machine learning techniques has the potential to shed new light on the complexity of metabolism [148, 149].

The scope of the present Chapter is to introduce to basic concepts underlying CBM and to describe a novel method to integrate it with SCalable UnBAlanced gene prioritization (Scuba) to prioritise candidate genes.

4.1 Constraint-based modelling of metabolism

In full generality, a metabolic system is composed by M metabolites that interact through chemical processes encoded in N reactions. The stoichiometric matrix $\mathbf{S}$ of the system is a $M \times N$ table containing the stoichiometric coefficients of all reactions for each metabolite. Entry S_{ij} is positive or negative if metabolite i is produced or consumed in reaction j, respectively. Let us indicate with $\mathbf{v} \in \mathbb{R}^N$ the vector of reaction rates - or reaction *fluxes* - of the system. The variation of metabolite concentrations $\mathbf{c}$ over time is described by the

following equation in matrix notation:

$$\mathbf{Sv} = \frac{d\mathbf{c}}{dt} \equiv \mathbf{b}\,. \tag{4.1}$$

At steady state, concentration of intermediate metabolites are constant: $d\mathbf{c}/dt = 0$. Therefore, the whole collection of reactions is represented by a set of linear equations, which can be expressed as follows:

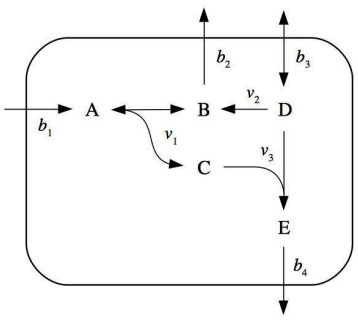
$$\mathbf{Sv} = 0\,. \tag{4.2}$$

This simple equation represents constraints given by the reaction network topology and stoichiometry. Since fluxes cannot assume arbitrarily large values, they are usually constrained also by a set of reasonable upper and lower bounds:

$$\mathbf{v}_{lb} \leq \mathbf{v} \leq \mathbf{v}_{ub}\,. \tag{4.3}$$

For instance, irreversible reactions are assigned a lower bound equal to 0, so that backwards flow is precluded.

To model the genetic dependency of metabolic reactions, they are linked to associated enzymes through logical rules. For reactions that require the action of enzymatic complexes to occur, the corresponding enzymes are bound by AND relationships. If instead there are multiple isozymes that catalyse the same reaction in parallel, they are linked by OR



$$v_1 : A \leftrightarrow B + C$$
$$v_2 : D \rightarrow B$$
$$v_3 : 2C + D \rightarrow E$$

$$\mathbf{S} = \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

Figure 4.1 | Example of a metabolic model and its stoichiometric matrix. Matrix **S** summarises the stoichiometry of the system, associating each row to a metabolite and each column to a reaction.

relationships. In this way it is possible to express the occurrence of a reaction in terms of the presence or absence of its required enzymes.

In a typical CBM problem, fluxes are the variables and have to be determined. Since the matrix **S** represents M equations for N variables and reactions are usually more than metabolites, the problem is under-determined - meaning that multiple solutions satisfy it. In order to determine biologically meaningful solutions, it is often necessary to further define the model by applying additional biological, physical or chemical constraints. They may for example account for enzyme capacity and promiscuity, spatial occupation, metabolite sequestration and multiple levels of gene, transcript and protein regulation. Analogously, artificial gene deletion or de-regulation can be simulated by acting on flux boundaries. For instance, a gene deletion can be introduced forcing its corresponding reactions to carry a null flux. This technique can be used to optimise the production of certain metabolites or to reproduce altered disease networks [150].

Particularly useful are constraints derived by experimental data, employed to build models reflecting directly observed biological conditions, like those in particular tissues or pathological states. Genome-wide transcriptional profiles are one of the most used data types to build such context-specific GSMMs. In the next section, approaches that serve this purpose are described.

4.2 Integration with transcriptional regulation

Enzyme availability level in GSMMs is usually quantified in terms of transcript abundance. The main motivation is the higher availability, quality and coverage of gene expression data as compared to protein expression. The level of agreement between the two quantities is matter of investigations and debate, given the contradicting results obtained so far [151–154]. Most recent studies show however that in human mRNA abundance is a good indicator of protein abundance, especially when averaging across populations, although differences exist depending on the considered genes and tissues.

Methods to integrate transcriptional regulation into metabolic models can traditionally be divided into two main categories [26]. The first one uses transcriptional profiles to build context-specific models, the second one integrates GSMMs with Boolean regulatory networks. Methods in the first category are descriptive rather than predictive, namely they are able to describe the regulation of metabolism in contexts with available omics data but they cannot make predictions on contexts lacking this data. Furthermore, these methods allow making predictions only on metabolic agents and reactions, as they do not explicitly use the regulatory processes underlying expression profiles. Conversely, methods in the second

category provide mechanistic control on gene regulation but they also require a long and tedious work to reconstruct the underlying network. Moreover, Boolean rules are in most of the cases simplistic compared to real regulatory processes. Only very recently, a new third class of methods emerged trying to combine expression profiles and transcriptional regulatory networks to create integrated metabolic-regulatory models [155].

In turn, methods belonging to the first class can be divided based on the kind of criterion used to contextualise the metabolic model [26]. Switch-based methods utilise a gene expression threshold to turn off reactions associated to lowly expressed genes and thereby prune the metabolic network. Valve-based methods map instead the transcriptional information on the model in a continuous fashion. In the next Subsection we describe a recent valve-based method used in our following study.

4.2.1 Gene set expression mapping

An effective way to integrate gene expression profiles on metabolic models is by means of gene set rules and logarithmic maps, as implemented in METRADE and, more recently, in a breast cancer study [27, 28]. The main assumption is that the original GSMM represents in some sense an unperturbed cellular condition. The transcriptional profile mapped on it represents the perturbation applied on the system. In this sense, the rationale is to consider an expression fold change for the condition of interest in comparison to a reference transcriptional level. Since metabolic reactions can be catalysed by multiple enzymes, gene expression fold changes are translated into effective gene set expression fold changes. To this end, logical OR/AND rules among genes involved in any reaction (Section 4.1), are converted into max/min operations as follows:

$$\begin{aligned} &\text{single gene:} && f(g_i) = f(g_i) \\ &\text{enzymatic complex:} && f(g_i \wedge g_j) = min\{f(g_i), f(g_j)\} \\ &\text{isozymes:} && f(g_i \vee g_j) = max\{f(g_i), f(g_j)\}, \end{aligned} \tag{4.4}$$

where $f(g_i)$ and $f(g_j)$ represent the expression fold changes for genes g_i and g_j. The fold change profile for all genes can be ideally calculated in two main ways. One possibility is to consider the variation between case and controls conditions, for instance corresponding to disease and healthy samples. Another possibility is to take as a reference the average expression in multiple conditions. For example, one may build a tissue-specific model by taking the fold change as compared to a set of tissues.

In several cases, reactions depend on the concentration of numerous gene sets, resulting in complex nested expressions. These rules can be recursively applied until a final effective

fold change f_k^{react}, associated to any reaction k, is obtained. This quantity is mapped on the flux bounds $\mathbf{v}_{lb}$ and $\mathbf{v}_{ub}$ by multiplying them by a factor defined as follows:

$$h(f_k^{react}) = \begin{cases} (1+\mu\,|log(f_k^{react})|)^{sgn(f_k^{react}-1)} & \text{if } f_k^{react} \in \mathbb{R}^+ \setminus \{1\} \\ 1 & \text{if } f_k^{react} = 1\,, \end{cases} \tag{4.5}$$

where $sgn(f_k^{react}-1) = (f_k^{react}-1)/|f_k^{react}-1|$. In this equation μ is a parameter representing the magnitude with which expression affects reaction rates. The larger μ and the more flux bounds are expanded (restricted) by a fold change larger (smaller) than 1. A variation of this approach consists in evaluating absolute gene expression values through Eq. 4.4 and then compute the fold change between effective reaction expressions, that can be mapped on model bounds via Eq. 4.5.

The choice of a logarithmic map from gene expression levels to reaction rates is supported by experimental evidence in various biological processes, both in bacteria and cancer cells [156, 157]. Independent findings suggest that protein synthesis rate is also faster at high mRNA concentration and slower at decreasing concentration [158]. Reasonably, this function smooths unrealistically large expression values while translating them into flux bounds. Moreover, around unitary values the fold change can be approximated by a linear function, matching ^{13}C-labelling experimental results [159].

4.3 Characterisation of the flux space

From a geometrical point of view, the set of solutions defined by any GSMM (Eq. 4.2,4.3) forms a polyhedron in the space of reaction fluxes [160]. As a simple illustrative example consider Figure 4.2. Red boundaries stemming from the origin represent stoichiometric constraints, while the bases of the flux cone are given by constraints on the single reactions. The region of space occupied by the polyhedron represents all the possible flux configurations that the system can assume. A legitimate question is which points in the solution space are in fact biologically meaningful. In principle, a cell might use only a limited subspace of the feasible region. As a matter of fact, bacteria live and proliferate on the edge of a Pareto optimality [27].

Depending on the questions one seeks to answer, there are various techniques that can be used to gain meaningful information on the shape or subsets of the allowed flux space. In the following Subsections, some of the main techniques are presented.

4.3.1 Flux balance analysis

The most popular approach to obtain biologically meaningful flux solutions is Flux Balance Analysis (FBA), which allows determining the flux vector $\mathbf{v}$ that yields maximal or minimal production of one or more target metabolites [138]. It is mathematically defined as an optimisation problem on a subset of target reaction fluxes. The formulation is the following:

$$\begin{aligned} &\min_{\mathbf{v}} / \max_{\mathbf{v}} \ \mathbf{w}^{\top} \cdot \mathbf{v} \\ &\text{subject to} \quad \mathbf{S} \cdot \mathbf{v} = 0 \\ &\mathbf{v}_{lb} \leq \mathbf{v} \leq \mathbf{v}_{ub} \,, \end{aligned} \tag{4.6}$$

where $\mathbf{w}$ is a real vector expressing the contribution of each reaction to the objective. Geometrically, FBA computes the extreme point coordinates along the axes of the specified objectives, as depicted in Figure 4.3.

Being formulated as a linear problem, this technique is very computationally efficient, however choosing a meaningful objective may be challenging. Usually, when no other obvious cellular objective is involved, maximisation of biomass is considered reasonable for bacteria under evolutionary pressure, but also for cancer cells under a proliferative regime

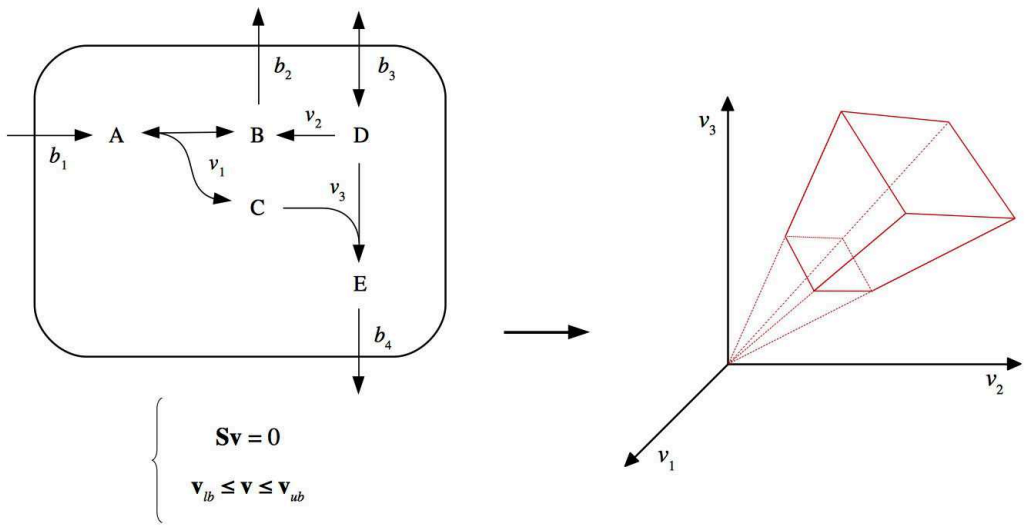


Figure 4.2 | Feasible flux space associated to a metabolic model. Graphical representation of the solution set associated to a simple metabolic model, shaped as a polyhedron in the reaction flux space.

[139, 161]. For other types of cells identifying the true objective is still a challenge, so biomass is commonly taken as a proxy.

4.3.2 Parsimonious enzyme usage flux balance analysis

Parsimonious enzyme usage Flux Balance Analysis (pFBA) is an extension of FBA that aims at computing the most economic flux distribution yielding a maximal biomass production. The underlying assumption is that, at exponential growth, cellular populations that most efficiently manage enzyme production are selected by evolutionary pressure [162].

This analysis is formulated as a bi-level optimisation in which first the biomass production rate is optimised via FBA and then the total sum of reaction fluxes is minimised. In order to achieve the second step, reversible reactions are split into two separate irreversible reactions and each irreversible reaction is constrained to carry a non-negative flux. The bi-level optimisation is defined as follows:

$$\begin{aligned} &\min_{\mathbf{v}_{irrev}} \sum_{i=1}^{N_{irrev}} \mathbf{v}_{irrev} \\ \text{subject to} \quad & v_{lb,biomass} = \max_{\mathbf{v}_{irrev}} v_{biomass}\,, \\ \text{subject to} \quad & \mathbf{S}_{irrev} \cdot \mathbf{v}_{irrev} = 0 \\ & 0 \leq \mathbf{v}_{irrev} \leq \mathbf{v}_{ub,irrev}\,. \end{aligned} \tag{4.7}$$

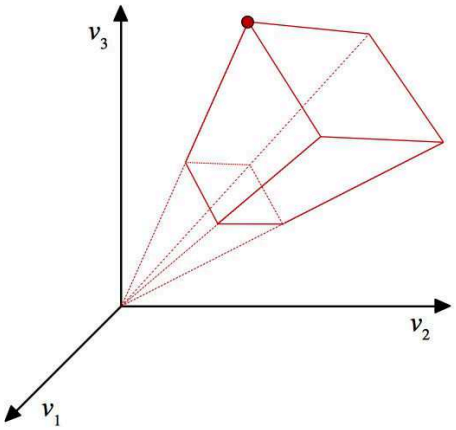


Figure 4.3 | Flux balance analysis identifies extreme points in the flux space. FBA identifies the extreme point in the flux space that maximises or minimises a given target flux or combination of fluxes.

Here $\mathbf{v}_{irrev}$ and N_{irrev} represent the flux vector of the irreversible model and its length, while $\mathbf{S}_{irrev}$ is the corresponding stoichiometric matrix. The lower bound for biomass production in the minimisation step $v_{lb,biomass}$ is defined by its optimal value resulting from the maximisation $v_{biomass}$. In this way, all fluxes are minimised except this one, guaranteeing optimal parsimonious growth.

However, minimising a sum is a not convex problem and may have multiple local optima. It is therefore possible to define a modified form of pFBA wherein the squared sum of reaction fluxes is minimised, which we denote here as euclidean parsimonious enzyme usage Flux Balance Analysis (epFBA). Analogously to before, the formulation is the following:

$$\begin{aligned} &\min_{\mathbf{v}_{irrev}} \sum_{i=1}^{N_{irrev}} \mathbf{v}_{irrev}^2 \\ \text{subject to} \quad & v_{lb,biomass} = \max_{\mathbf{v}_{irrev}} v_{biomass}\,, \\ \text{subject to} \quad & \mathbf{S}_{irrev} \cdot \mathbf{v}_{irrev} = 0 \\ & 0 \leq \mathbf{v}_{irrev} \leq \mathbf{v}_{ub,irrev}\,. \end{aligned} \tag{4.8}$$

Except for the second step objective, the procedure remains identical to pFBA.

4.3.3 Flux distribution profiling

When less strict assumptions are made, it can be of interest to map the whole flux solution space. In such a situation, Monte Carlo sampling can be used to determine the flux probability distribution without prior knowledge [163, 164]. The main drawback of this method is that uniform sampling is guaranteed only in the asymptotic limit, so time-consuming calculations are required. Indeed, the GSMM solution space is typically much more elongated along a few directions, fact that challenges current sampling algorithms.

A different recently proposed method allows analytically computing the flux distribution in metabolic networks [165]. It is based on a Bayesian approach called expectation propagation algorithm and here we denote it as Metabolic Expectation Propagation (MEP). Notably, it can efficiently profile flux distributions for all reactions even of large systems - such as the human metabolism - under some assumptions.

In particular, MEP implements an iterative procedure to compute the multivariate flux distribution $Q(\mathbf{v}|\mathbf{b})$ by defining a quadratic energy function $E(\mathbf{v})$ whose minimum(s) correspond to the stoichiometric constraints of Eq. 4.1:

$$E(\mathbf{v}) = \frac{1}{2}(\mathbf{S}\mathbf{v} - \mathbf{b})^T(\mathbf{S}\mathbf{v} - \mathbf{b})\,, \tag{4.9}$$

Each iteration step includes a sequential update of the following posterior multivariate flux distribution:

$$Q^{(i)}(\mathbf{v}|\mathbf{b}) = \frac{1}{Z_{Q^{(i)}}} e^{-\beta E(\mathbf{v})} \psi_i(v_i) \prod_{j \neq i} \phi_j(v_j), \tag{4.10}$$

where $\psi_i(v_i)$ and $\phi_j(v_j)$ represent the priors for reactions i and j respectively. $Z_{Q^{(i)}}$ is a normalisation factor defined as follows:

$$Z_{Q^{(i)}} = \int d^N \mathbf{v}\, e^{-\beta E(\mathbf{v})} \psi_i(v_i) \prod_{j \neq i} \phi_j(v_j). \tag{4.11}$$

In these equations, distributions $\phi_j(v_j)$ pertaining to all reactions but the i^{th}, are approximated by normal distributions. The idea is to treat exactly only the distribution $\psi_i(v_i)$ at each step, thereby making calculations tractable. The update of the posterior distribution is performed via matching its mean and variance to those of a fully Gaussian-approximated multivariate distribution Q:

$$\begin{cases} \langle v_i \rangle_{Q^{(i)}} = \langle v_i \rangle_Q \\ \langle v_i^2 \rangle_{Q^{(i)}} = \langle v_i^2 \rangle_Q . \end{cases} \tag{4.12}$$

It was shown that for an increasing number of iterations, mean and variance of flux distributions computed through MEP converge to those obtained by sampling, but in only a fraction of execution time.

4.4 Fluxome: an integrative omic for gene prioritisation

The integration of metabolic CBM with machine learning lies on two main key ideas. The first is that genetic perturbations propagate in a non-linear fashion through metabolic networks and can assume informative patterns on a metabolic level that are useful to identify disease genes. The second is that GSMMs can be both an analytical framework to represent biological systems and generators of information to be mined. In other words, flux solutions obtained by a GSMM can be treated just like other numerical data and analysed via learning algorithms.

We propose a CBM-based prioritisation of candidate disease genes as follows:

i. **Build a condition-specific model.** Create a metabolic model relevant to the disorder of interest, by mapping experimental gene expression profiles on a general purpose GSMM. In particular, a possible option is creating a models specific for the interested tissue or cell type.

ii. **Build genetic perturbation-specific models.** In turn, perform *in silico* up/down-regulation of each known or candidate disease gene in the condition-specific model. In this way a different perturbed model is associated to each de-regulation.

iii. **Characterise the flux space perturbation-specific models.** Extract information on the feasible flux space of each perturbed model, in order to estimate flux alterations associated to each genetic perturbations.

iv. **Prioritise via machine learning.** Predict disease genes using the information acquired in the third step, through PU learning.

In this procedure, the prioritisation approach is more similar to a guilt-by-rewiring than to a Guilt-By-Association (GBA) (Section 1.3), mimicking functional knock-out screenings on human cells [35, 166].

Although this pipeline is valid in full generality and different methods can be chosen for each step, we advance a more precise proposal. To build condition-specific models, we propose to adopt the gene set expression mapping described in Section 4.2.1. Equations 4.4 and 4.5 allow varying reaction bounds in a continuous way maintaining a constant number of reactions, which ultimately translates in a constant amount of flux features. In order to characterise the metabolic state of perturbation-specific models, we propose to utilise MEP, as it is currently the most efficient technique to obtain maximal information about the entire solution space. Finally, we propose to infer putative gene-disease associations by means of Scuba.

As discussed in a recent paper, GSMMs provide a complementary perspective compared to high-throughput data like Protein-Protein Interactions (PPIs) [167]. Such data is experimentally generated - hence require no prior knowledge - and can in principle span the entire genome/transcriptome/proteome, although experimental biases exist. On the other hand, it is also prone to contain false-positive interactions and can be superficial or ambiguous on its biological meaning. In contrast, GSMMs are often limited to metabolic networks but are highly curated and provide mechanistic description of biological processes, linking together genes, enzymes, metabolites and reactions. Moreover, as compared to annotations, they can more precisely describe the functional role of genes as they provide a direct representation of biochemical processes without an ontology or abstract semantics. Therefore, flux analysis is expected to provide complementary information and hopefully support disease gene identification.

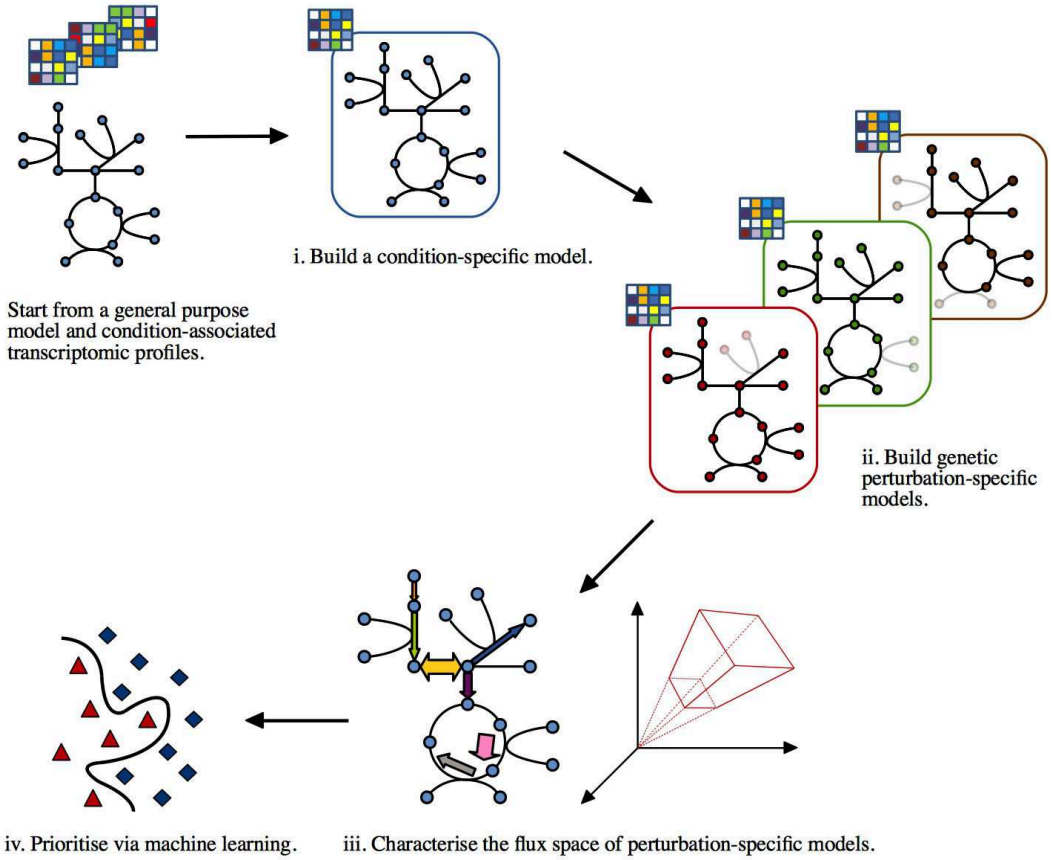


Figure 4.4 | Work-flow to combine constraint-based modelling and machine learning. The condition-specific model represents the disease phenotype of interest and is build from the combination of a general purpose model and disease or tissue transcriptional profiles. The characterisation of the flux space serves the production of gene-specific information, mined via machine learning prioritisation methods.

Chapter 5

Predicting disease genes via integrative *in silico* metabolic flux profiling

Constraint-based modelling of metabolism is employed as an information source for prioritisation of oncogenes and tumour suppressor genes. Predictions based on metabolic flux profiles are benchmarked against those obtained starting from transcriptomics data and pathway annotations. Despite a highly variable effectiveness across cancer types, a general complementarity of predictions is observed.

In this Chapter we investigate the integration of mathematical modelling with experimental omics data and gene annotations, aimed at improving computational gene prioritisation. As described in the previous Chapter, we identified Constraint-Based Modelling (CBM) as a promising modelling methodology, due to its flexibility in integrating omics information at a genome scale. We thus applied the method presented in Section 4.4 to investigate the role of metabolism in propagating genetic perturbations. Given the recent breakthroughs in understanding the metabolic nature of cancer, we assessed the benefit of integrating *in silico* metabolic flux information for cross-cancer gene prioritisation [168].

5.1 Building of tumour-specific metabolic models

As a first step toward the prioritisation of candidate cancer genes, we constructed tumour-specific metabolic models by mapping experimental transcriptional data onto a human Genome-Scale Metabolic Model (GSMM), Recon 2.2 [29]. To this end, we used the gene set expression mapping presented in Section 4.2.1. The strength of the logarithmic factor in Eq. 4.5 is controlled by a parameter μ which modulates the impact of gene expression

on reaction bounds. We therefore tuned this parameter to optimise the resulting models, maximising the agreement with experimental data.

In the following, we present the process of building and validating tumour-specific GSMMs via two different strategies. In the first case, we used gene expression data from a popular cancer cell line panel. In the second case, we employed data collected directly from cancer patients.

5.1.1 Cell lines models

In this phase, we resorted to experimental data collected from the National Cancer Institute 60 human tumour cell line (NCI60) panel [169]. We extracted from it gene expression profiles and proliferation rates for 57 cell lines associated to nine different cancers, already pre-processed in a previous study [170]. The proliferation rate is expressed as the inverse of cell number doubling time.

In order to assess the reliability of the resulting models, we performed a sensitivity analysis on parameter μ, in terms of the Pearson Correlation Coefficient (PCC) between predicted cellular growth and measured proliferation rate of the NCI60 panel. In other words, we mapped on Recon the expression profiles and for each cell line-specific model we performed Flux Balance Analysis (FBA) by setting the biomass reaction as objective. By assumption, this value should have the same trend as the proliferation rate, so we calculated

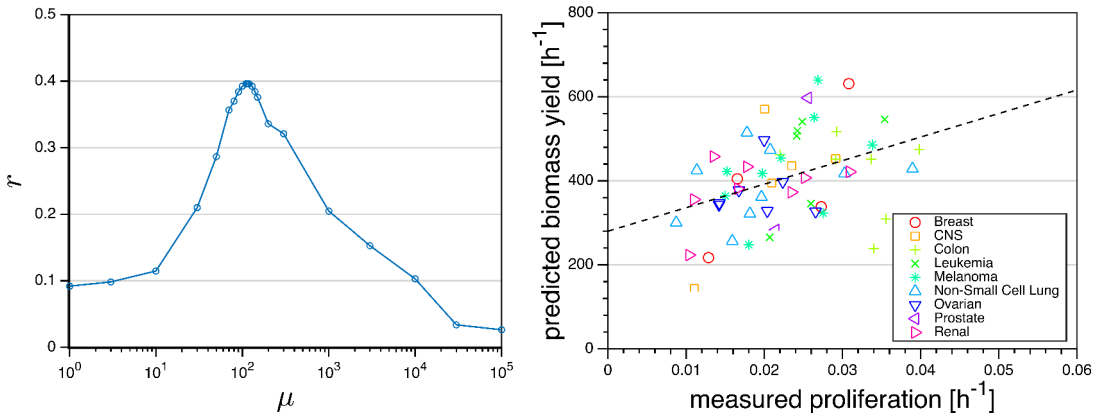


Figure 5.1 | Sensitivity analysis on cell lines gene expression mapping. **(a)** PCC r between predicted biomass yield and measured cellular proliferation, as a function of the gene expression mapping parameter μ. **(b)** Predicted biomass yield for each NCI60 cell line as compared to its corresponding measured proliferation, at the maximal r value.

the correlation between the proliferation and the optimal flux through the biomass reaction. We repeated this procedure varying the value of μ over a span of different orders of magnitude. As a result, we obtained a correlation peak around $\mu = 114$, where the PCC is $r = 0.41$, p-value$= 1.56 \cdot 10^{-3}$ (see Figure 5.1a). In Figure 5.1b it is possible to visually appreciate the correlation between predicted biomass yield and measured proliferation over all the 57 cell lines. These results provide us an indication of the good reliability of constructed models. We performed this analysis also restricting to each cancer group, however we did not obtain statistically significant correlations - except in one case. We therefore assumed $\mu = 114$ as the optimal mapping value for all tumours in the following analyses.

Next, we further validated the constructed models by focusing on the whole set of reaction fluxes. We thus studied the PCC between proliferation and flux values for all reactions in the

Pathway	**Number of correlated reactions**	**Fraction of correlated reactions (%)**	**Supporting literature**
Aminosugar metabolism	2	6.45	
Butanoate metabolism	1	33.33	
Cholesterol metabolism	13	22.81	[171]
Citric acid cycle	1	5.26	
Exchange/demand reaction	51	6.93	[172]
Fatty acid oxidation	8	0.99	[173–175]
Fatty acid synthesis	1	0.85	[176]
Folate metabolism	3	5.08	[177]
Glutamate metabolism	1	6.67	[178]
Glycerophospholipid metabolism	3	4.55	[179]
Glycine, serine, alanine and threonine metabolism	1	2.70	[180]
Miscellaneous	4	4.65	
NAD metabolism	2	8.00	[181, 182]
Purine catabolism	2	5.56	
Pyrimidine catabolism	1	2.86	
Sphingolipid metabolism	3	3.61	[183]
Squalene and cholesterol synthesis	2	33.33	[171]
Starch and sucrose metabolism	1	3.13	[184, 185]
Transport, endoplasmic reticular	4	2.60	
Transport, extracellular	99	6.73	[172]
Transport, lysosomal	1	0.94	
Transport, mitochondrial	8	2.78	[186]
Transport, peroxisomal	1	0.85	
Other	39	4.57	

Table 5.1 | Overview of metabolic reactions whose predicted fluxes correlate with measured cellular proliferation. For each pathway, number and percentage of reactions significantly correlated to measured cellular proliferation at a 1% threshold. Several of these pathways are known to be involved in tumour initiation or evolution, as reported in the corresponding literature.

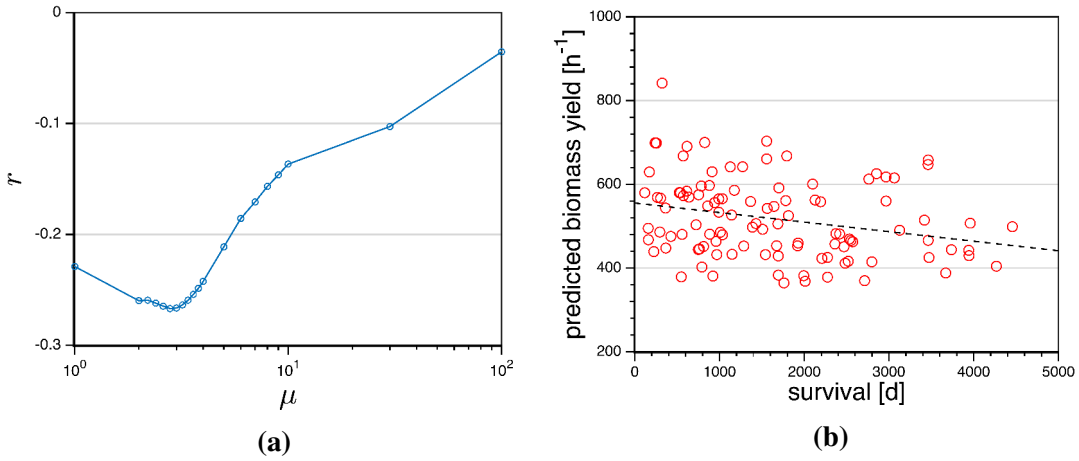


Figure 5.2 | Sensitivity analysis on breast cancer patients gene expression mapping. **(a)** PCC r between predicted biomass yield and patient survival time span, as a function of the gene expression mapping parameter μ. **(b)** Predicted biomass yield for each breast cancer patient as compared to the corresponding survival time span, at the maximal r value.

models, calculated through euclidean parsimonious enzyme usage Flux Balance Analysis (epFBA) (Section 4.3.2). The rationale is that tumour cells need to optimise the use of resources to proliferate in the competitive environment with normal cells. We thus expect that epFBA better captures the flux distribution in cancer cells than normal FBA. We took p-value ≤ 0.01 as a threshold for designating correlated reactions and we considered both positive and negative correlations. In Table 5.1 we show absolute and relative number of correlating reactions for each pathway. Pathway information was obtained directly from Recon, available in the *subSystem* field of the Matlab model [29]. The full list of 252 correlating reactions can be found in Table B.1.

We can observe that reaction rates obtained by epFBA correlate with cellular proliferation in a number of cancer-associated pathways, supporting the reliability of our models. In particular, our results suggest that several identified reactions are involved in extracellular transport, which is consistent with the known metabolic interactions between tumour cells and their micro-environment [172]. Moreover, a notable portion of the cholesterol metabolism pathways emerges as correlated to proliferation, supported by epidemiologic, Next-Generation Sequencing (NGS) and preclinical data that suggest a dependence of patient survival on alterations in cholesterol homeostasis [171].

Pathway	Number of correlated reactions	Fraction of correlated reactions (%)	Supporting literature
Arginine and Proline Metabolism	2	5.13	
Bile acid synthesis	8	6.40	[187, 188]
Cholesterol metabolism	11	19.30	[171]
Eicosanoid metabolism	1	0.40	
Exchange/demand reaction	11	1.49	[172]
Folate metabolism	2	3.39	[177]
Fructose and mannose metabolism	1	4.00	
Glycerophospholipid metabolism	1	1.52	[179]
Glycine, serine, alanine and threonine metabolism	3	8.11	[180]
Glycolysis/gluconeogenesis	7	17.50	[189]
Nucleotide interconversion	1	0.56	
Pentose phosphate pathway	2	5.71	
Pyrimidine synthesis	1	5.26	
Pyruvate metabolism	1	3.33	
Sphingolipid metabolism	3	3.61	[183, 190]
Squalene and cholesterol synthesis	1	16.67	[171]
Transport, endoplasmic reticular	4	2.60	
Transport, extracellular	15	1.02	[172]
Transport, lysosomal	1	0.94	
Transport, mitochondrial	3	1.04	[186]
Transport, nuclear	1	1.56	
Transport, peroxisomal	1	0.85	
Other	13	1.52	

Table 5.2 | Pathways correlated to patient survival in breast cancer models. For each pathway, number and percentage of reactions significantly correlated to breast cancer patients survival at a 1% threshold. Several of these pathways are known to be involved in tumour initiation or evolution, as reported in the corresponding literature.

5.1.2 Patient samples models

A limitation of the analysis above is that it relies on data derived from cell cultures, which only approximate what occurs in a living human body. From this point of view, acquisition of patient samples can be a valuable strategy for better modelling cancer metabolism. However, even though patient specimens are beginning to be routinely collected, it is difficult to precisely track tumour proliferation. To evaluate patient-specific GSMMs, patient survival can be used instead as a reference. In fact, death of oncological patients is caused by cancer invasion and metastases accumulation, so it can be assumed that average life span is inversely proportional to tumour cells growth rate.

Under these assumptions, we repeated the analysis of the previous Section by employing patients data gathered at a Genome Data Analysis Center (GDAC) repository [191]. To

maintain the coherency of the analysis across different cancer types, we used Level 3 normalised gene expression and clinical data available as of January 28th, 2016 standard data run. We filtered tumour samples in a way to leave only those possessing both RNA sequencing and patient survival information. Survival is expressed as the number of days to death from the acquisition of samples. In this case, setting flux constraints to cancer GSMMs was performed on the basis of gene expression fold change between tumour and normal samples.

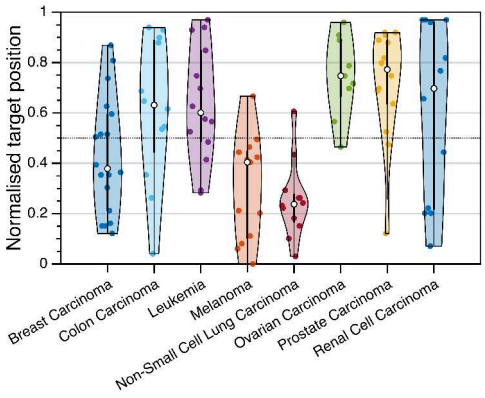
We performed sensitivity analyses for various tumour types separately, but only for breast cancer we obtained a significant PCC: $r = -0.27$, p-value$= 5.03 \cdot 10^{-3}$. As expected, the correlation is lower than considering proliferation, as survival represents an indirect estimate of cellular proliferation. Again, we scrutinised correlated reactions and identified several pathways known to be involved in cancer evolution (Table 5.2). This time the bile acid synthesis pathway emerges, which was found to influence growth in breast cancer cells [188]. For the complete list of correlated reactions we refer the reader to Table B.2.

5.2 Cross-cancer gene prioritisation

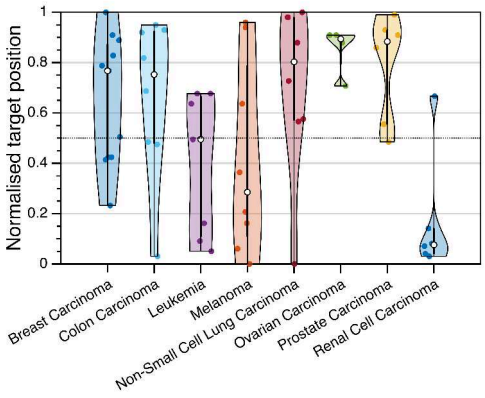
In this Section we focus on the prediction of metabolic oncogenes and tumour suppressor genes. The central assumption is that we should be able to predict cancer-related genes starting from metabolic perturbations caused by their upstream genetic de-regulation, as explained in Section 4.4.

We performed separate prioritisations for each cancer type covered by the NCI60 panel, with the only exclusion of central nervous system tumours because of the higher cell type heterogeneity in corresponding lines. For each cancer type, we generated a model using the gene expression mapping presented in Section 4.2.1. In this way we are able to simulate different cellular configurations and embed all artificial gene de-regulations in a different cellular context.

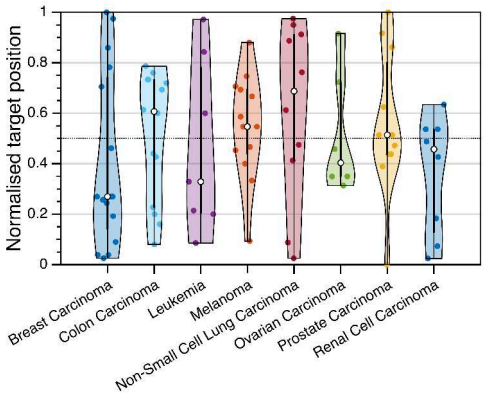
Moreover, we separately focused on the prioritisation of cancer genes in three different scenarios: full deletion of tumour suppressors, knock-down of oncogenes and knock-up of oncogenes. To simulate gene up-regulation and down-regulation at a single gene resolution, we employed again the mapping of Equations 4.4 and 4.5. In this case, instead of multiple experimental gene expression profiles, we consider a single artificial differential profile representing the expression fold change. Therefore, to simulate single gene up-regulation we take a fold change vector where the target gene is assigned a value greater than one and all other values are unitary. Conversely, we assign a fold change smaller than one to any target gene to be down-regulated, giving to all other genes an unitary fold change. This procedure



(a)



(b)



(c)

Figure 5.3 | AUC distributions for cancer genes prediction from cell line models. AUC distributions obtained by LOOCV for cancer types covered by the NCI60 panel and the prioritisation of **(a)** knocked-out tumour suppressor genes; **(b)** knocked-down oncogenes; **(c)** knocked-up oncogenes. Performances are highly variable depending on cancer type and de-regulation strategy. In particular, up-regulation of oncogenes fails in most cases on average.

allows us to up-regulate a broader set of genes as compared to the COBRA toolbox functions, which are based on logical OR/AND rules [22]. Specifically, we took as reasonable fold change factors 0.5 to simulate down-regulation and 2 for up-regulation.

We gathered reliable collections of oncogenes and tumour suppressor genes from recent published databases, TSgene and ONgene [192, 193]. Among them, 31 unique oncogenes and 57 unique tumour suppressor genes are included in Recon. To obtain cancer-specific oncogenes and tumour suppressors, we crossed these two lists with DisGeNet annotations [194]. We took a different set of 99 random genes as controls for every combination of cancer tissue, gene type and regulation scenario.

We generated a feature vector for each cancer and control gene using the Metabolic Expectation Propagation (MEP) (Section 4.3.3) [165]. Metabolic perturbations were therefore described in terms of means and variances of flux distributions through all reactions. From MEP we obtained mean and variance of bounded flux distributions for all reactions in the model, as well as mean and variance of their corresponding unbounded distributions. Due

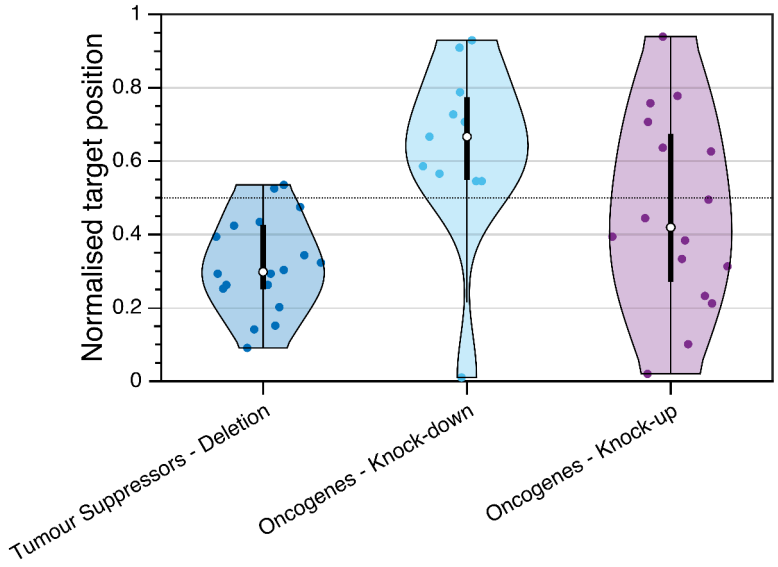


Figure 5.4 | AUC distributions for breast cancer genes prediction from patient models. AUC distributions obtained by LOOCV for down-regulated tumour suppressors and oncogenes and for up-regulated oncogenes in breast cancer patient models. In this scenario metabolic models do not provide useful information overall.

to the large number of simulations involved in our study, we limited the maximum number of iterations of the algorithm to 200. Ideally, allowing a more precise refinement of flux distributions could further increase prediction performance. Other parameters were set as suggested in the corresponding publication: $\beta = 10^9$, flux variance range $[10^{-50}, 10^{50}]$, damping coefficient $d = 0.9$. We aggregated flux information for reactions belonging to the same pathways by taking average and variance at a pathway level, in order to maintain a contained number of features and thus limit over-fitting. Association between pathways and reactions was taken from the *subSystem* field of Recon as before. We next input these pathway-based features into the machine learning step, where we employed SCalable UnBAlanced gene prioritization (Scuba) to rank genes on the basis of the cancer-specific metabolic perturbations triggered by their de-regulation. A Gaussian kernel with $\sigma = 0.5, 1, 2$ was employed to measure the similarity among flux profiles (Section 2.2.1).

We evaluated the whole pipeline by Leave-One-Out Cross Validation (LOOCV), measuring prediction performance in terms of Area Under the receiver-operating-characteristic Curve (AUC) for ranked genes. Single gene AUC distributions are visible in Figure 5.3. Moreover, we then performed the same evaluation by resorting to breast cancer patient models. These last results are shown in Figure 5.4.

In Chapter 3 we saw how network-based prioritisation is strongly driven by node degree. We thus decided to investigate whether similar biases exist in the context of CBM. In this case, we quantified the centrality of a gene as the number of its associated reactions in the model and we measured the Spearman Correlation Coefficient (SCC) ρ between centrality and AUC. As it is visible in Figure 5.5, no correlation exists for tumour suppressors, whereas a significant ρ can be detected for oncogenes. The bias is however far lower than for graph-like data, consistently with the concept that also non-central genes can influence cellular phenotype toward recognisable pathological states.

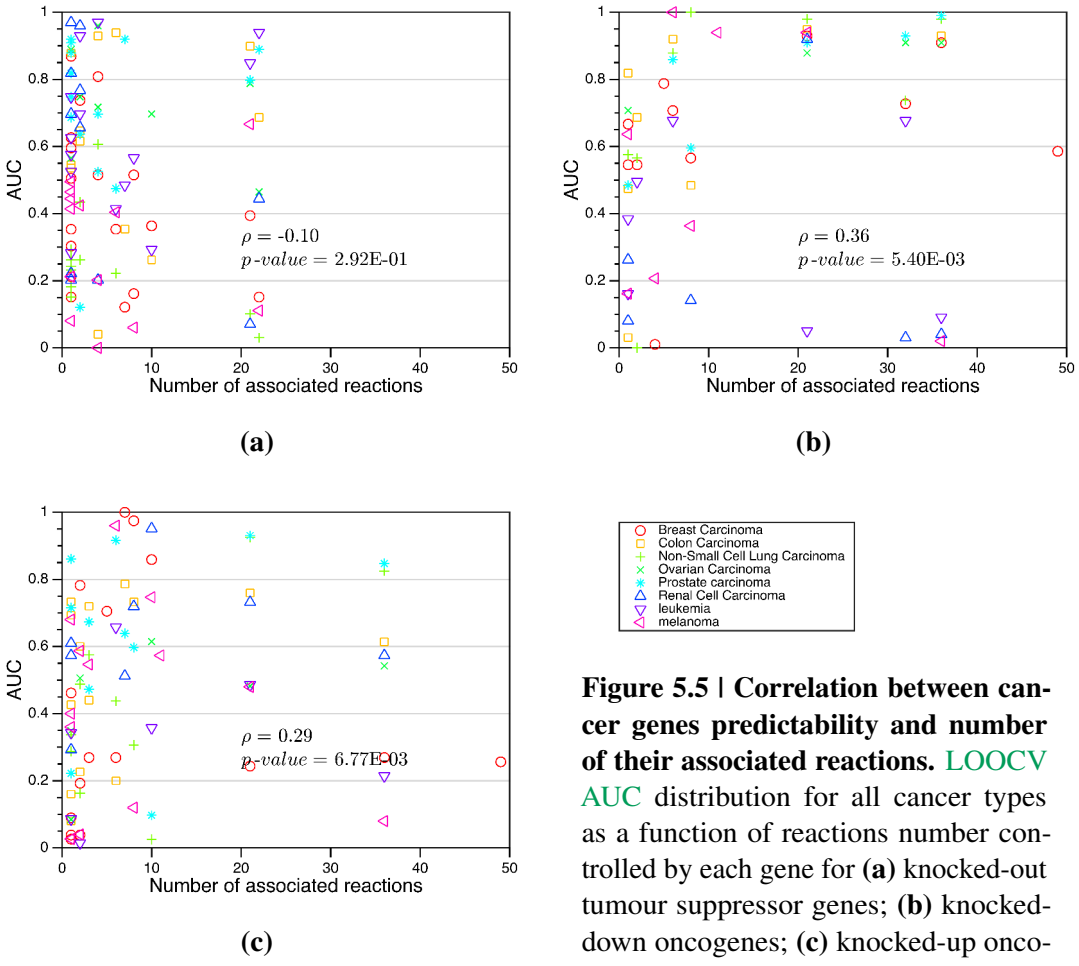


Figure 5.5 | Correlation between cancer genes predictability and number of their associated reactions. LOOCV AUC distribution for all cancer types as a function of reactions number controlled by each gene for **(a)** knocked-out tumour suppressor genes; **(b)** knocked-down oncogenes; **(c)** knocked-up oncogenes.

5.3 Integration of the fluxome with other omics

Upon verifying that *in silico* metabolic flux profiling can correctly suggest gene-cancer relationships, we turned to evaluating its effectiveness as compared to that of some conventional omics data. We performed LOOCV as before, using the same sets of known cancer genes and candidates while varying the input data sources. More in detail, we separately tested gene expression profiles from the NCI60 panel and two different pathway annotation datasets. The first one corresponds to Recon internal subdivision, composed of 98 metabolic pathways (excluding the "Miscellaneous "pathway and not assigned reactions). The second dataset is the most comprehensive pathway annotation resource to date, Consensus Path DataBase (CPDB), which contains information from a number of orthogonal sources on all known biological pathways [75]. All these three datasets were transformed by a Gaussian kernel with $\sigma = 0.5, 1, 2$ as for flux data.

Comparing prioritisation performance at a single gene resolution, we can notice total absence of correlation in all cases (Figure 5.6). Although there are differences in the average performance of the single data types, they remain rather complementary, with AUC values generally lying far from the diagonal. Such a distribution indicates that genes ranked at the bottom of candidate lists can be often ranked at the top by resorting to another data type. From this point of view, pathway-averaged metabolic flux profiles provide strikingly complementary results as compared to information on gene involvement in the same pathways (Figures 5.6a, 5.6b).

Next, we evaluated prediction performance for single cancer types, focussing on the differences in data integration approaches here considered, namely Multiple Kernel Learning (MKL) as implemented by Scuba and CBM. The latter can be employed in this case to integrate transcriptional data, whereas the former can process indiscriminately both expression and pathway annotations. Additionally, these two approaches can be combined by means of a multi-staged integration: first, building an integrated transcriptional-metabolic model via CBM; second, combining the information extracted from such model with other data sources via MKL.

In Figures 5.7 - 5.10, AUC distributions relative to all data and integration framework combinations are shown. First, it can be noticed that flux information tends to be more informative than gene expression in tumour suppressor prediction, while for oncogenes the contrary occurs. This is consistent with the biological model where tumour suppressor genes need to be compromised on both alleles for tumour to progress. Differently, oncogenes de-regulation depends from case to case and can be more difficult to pinpoint. As a matter of fact, changes in expression levels may assume a continuous spectrum of values, while we implemented fixed fold changes over all gene sets and tissues.

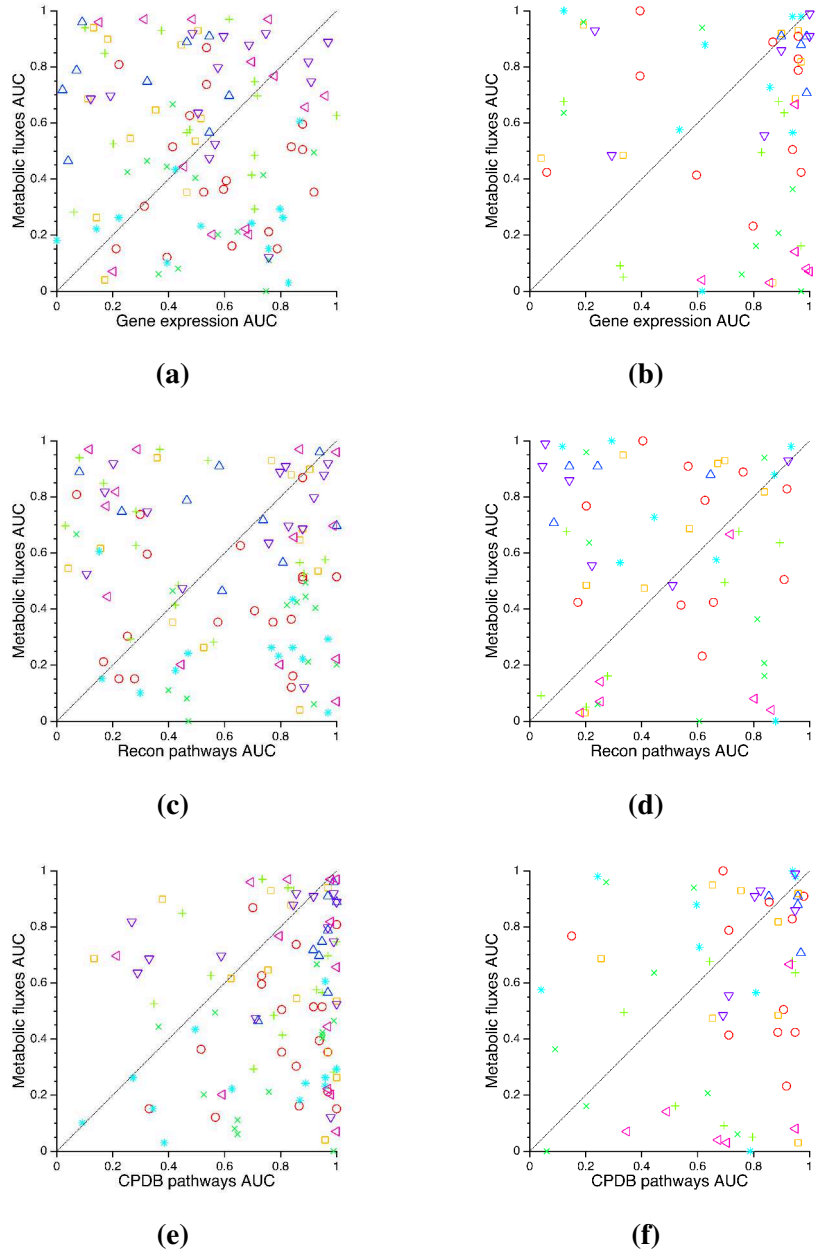


Figure 5.6 | Correlation between cancer genes predictions from different data sources. Comparison of single gene AUCs in prioritisation of knocked-down tumour suppressors (boxes **(a)**, **(c)** and **(e)**) and knocked-down oncogenes (boxes **(b)**, **(d)** and **(f)**).

In contrast, metabolic pathway annotations result much informative for tumour suppressors, but less for oncogenes. In fact, they are outperformed by metabolic flux data only in predicting leukemia and ovarian carcinoma tumour suppressors, while it occurs in five cases for oncogenes. However, considering all pathways in CPDB, AUCs sway more consistently in favour of annotations, evidencing that taking into account all biological domains at a coarse grain scale is more informative than only the metabolic domain at a more precise scale. Cancer is indeed implicated also at a signalling and regulatory level besides metabolism.

As regards the integration strategies, CBM outperforms all single data sources three times in tumour suppressor prediction and only once in oncogene prediction. Scuba achieves higher median AUC only twice for oncogenes instead, whereas in all other cases single datasets provide better predictions. In parallel, GSMM-based integration tends to win over Scuba in tumour suppressor prioritisation, whereas it is less effective in oncogene prioritisation. Again, this reflects shortcomings in our modelling approach mentioned above. In general, integration via CBM is more successful in tasks where Recon is more informative, that is where biological assumptions can be modelled more precisely. Conversely, where gene expression represents the higher degree of precision, Scuba is a better choice to combine input data.

It must be mentioned that AUC distributions shown are only rarely significantly different as assessed by Wilcoxon rank sum tests. Due to the low number of cancer genes involved, significance is achieved only for differences in average AUC of at least $\sim$0.3. The relatively low improvement by Scuba suggests that it is often unable to weight properly the input information.

To conclude, we can ascertain that these results promisingly show the advantage of considering metabolic rearrangements in predicting cancer genes. Success is however strongly dependent on the tumour type and on the assumptions made to simulate genetic alterations, besides the quality of the metabolic model.

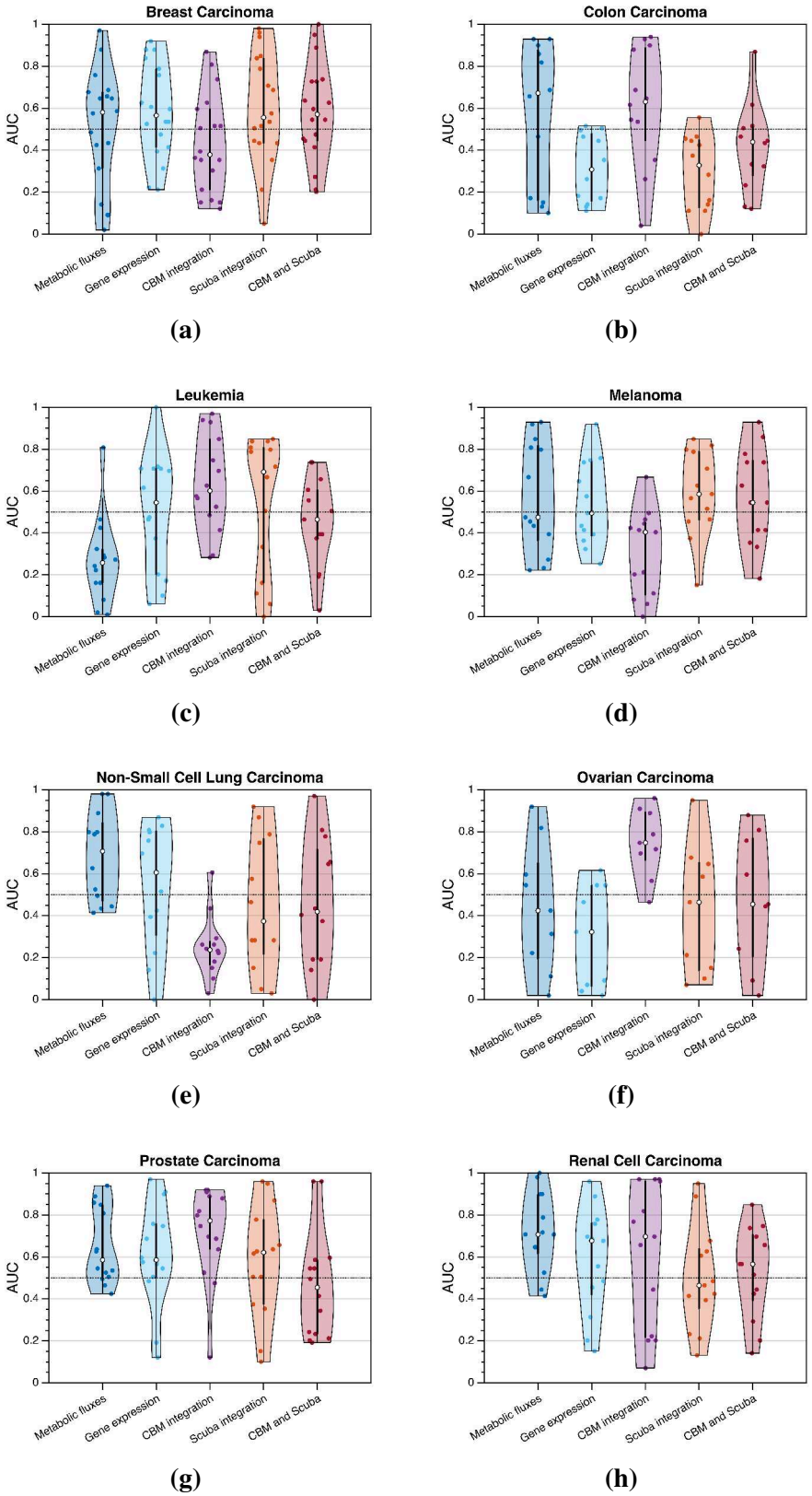


Figure 5.7 | Comparison between flux-based and gene expression-based prioritisation - Tumour suppressor genes. Comparison of tumour suppressors AUC distributions among single data sources and two alternative integrations approaches: MKL and CBM. In this case, we consider metabolic flux and gene expression data.

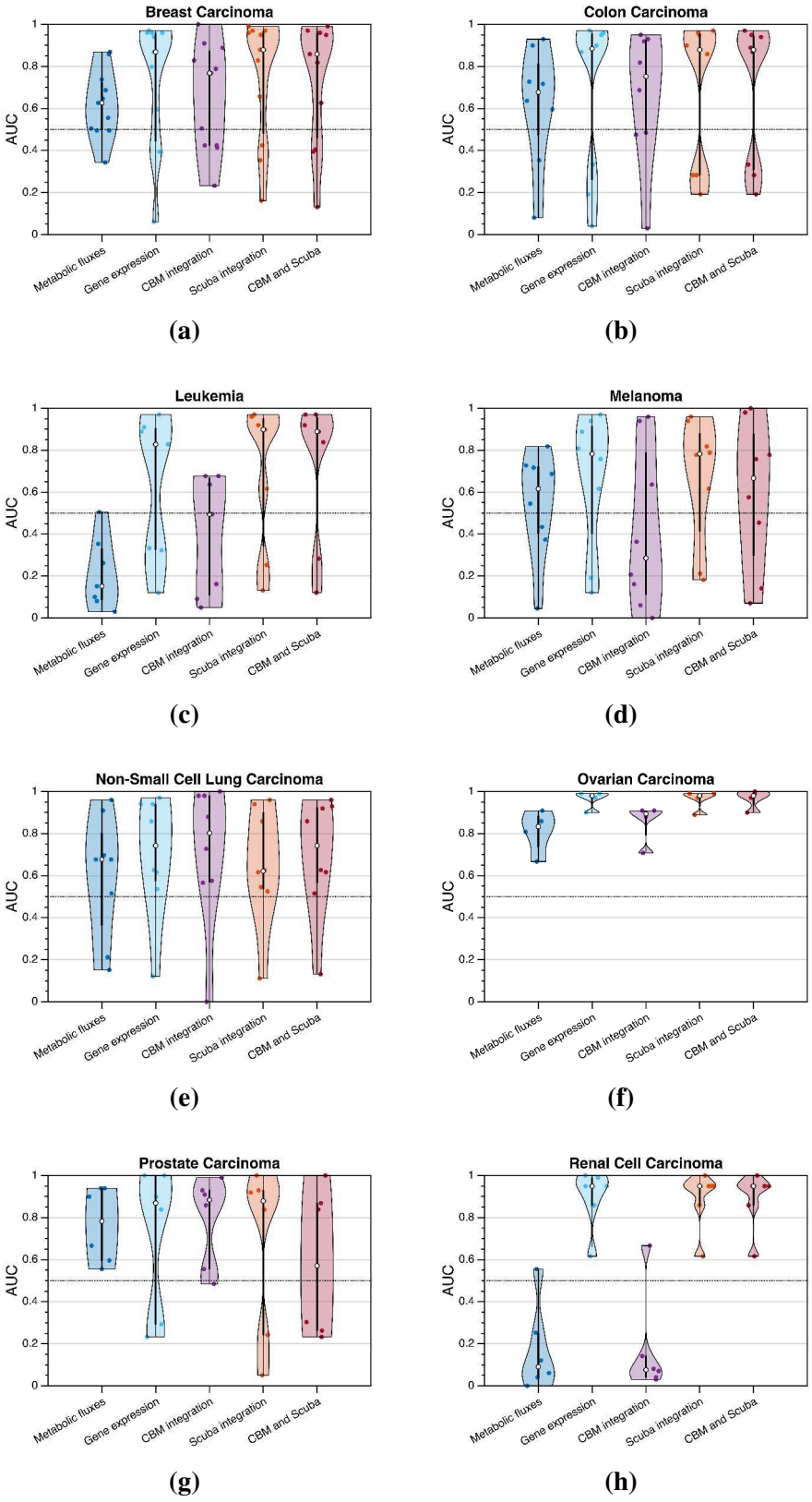


Figure 5.8 | Comparison between flux-based and gene expression-based prioritisation - Oncogenes. Comparison of oncogenes AUC distributions among single data sources and two alternative integrations approaches: MKL and CBM. In this case, we consider metabolic flux and gene expression data.

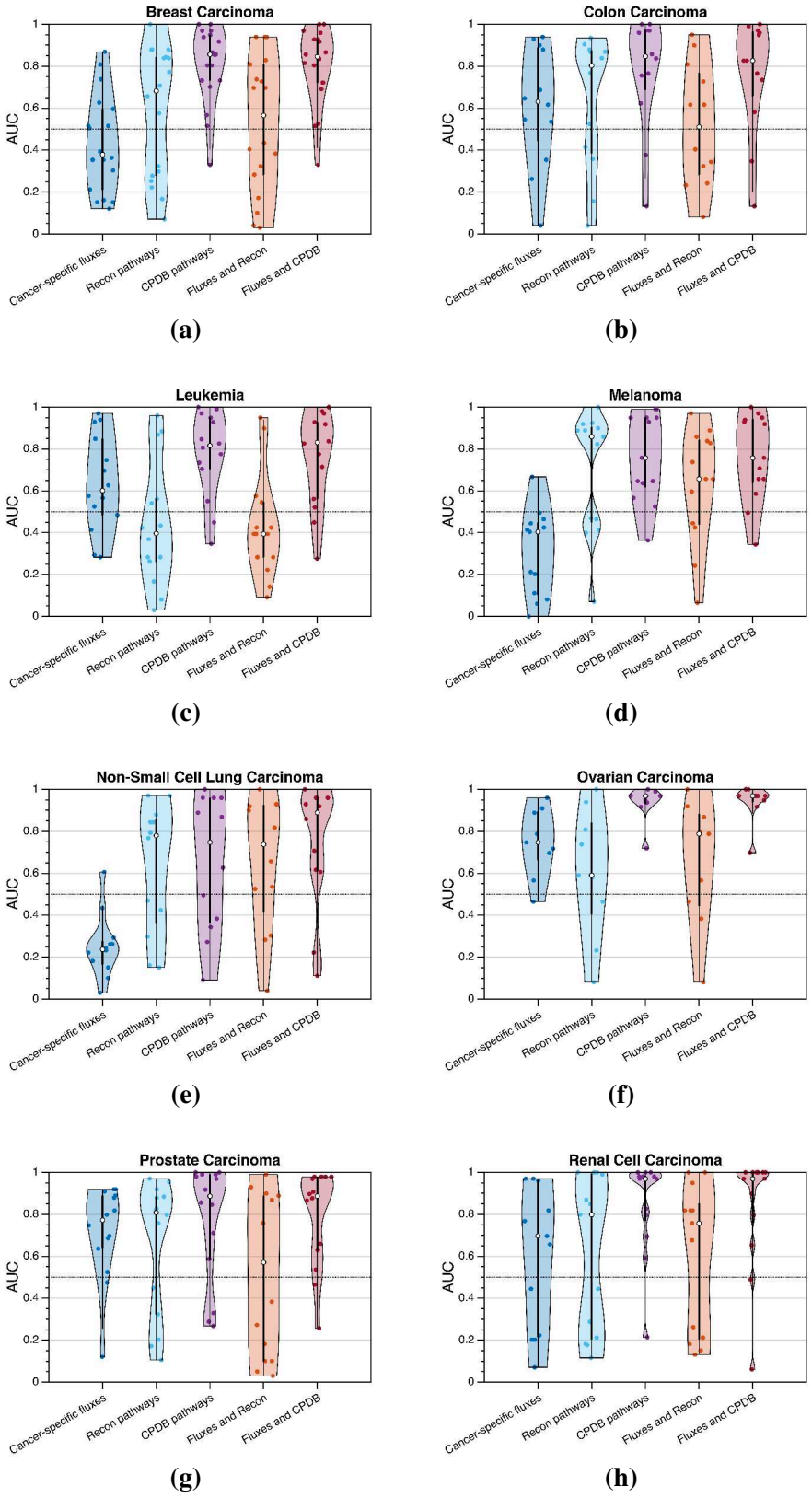


Figure 5.9 | Comparison between flux-based and pathway annotation-based prioritisation - Tumour suppressor genes. Comparison of tumour suppressors AUC distributions among single data sources and two alternative integrations approaches: MKL and CBM. In this case, we consider metabolic flux and pathway annotation data.

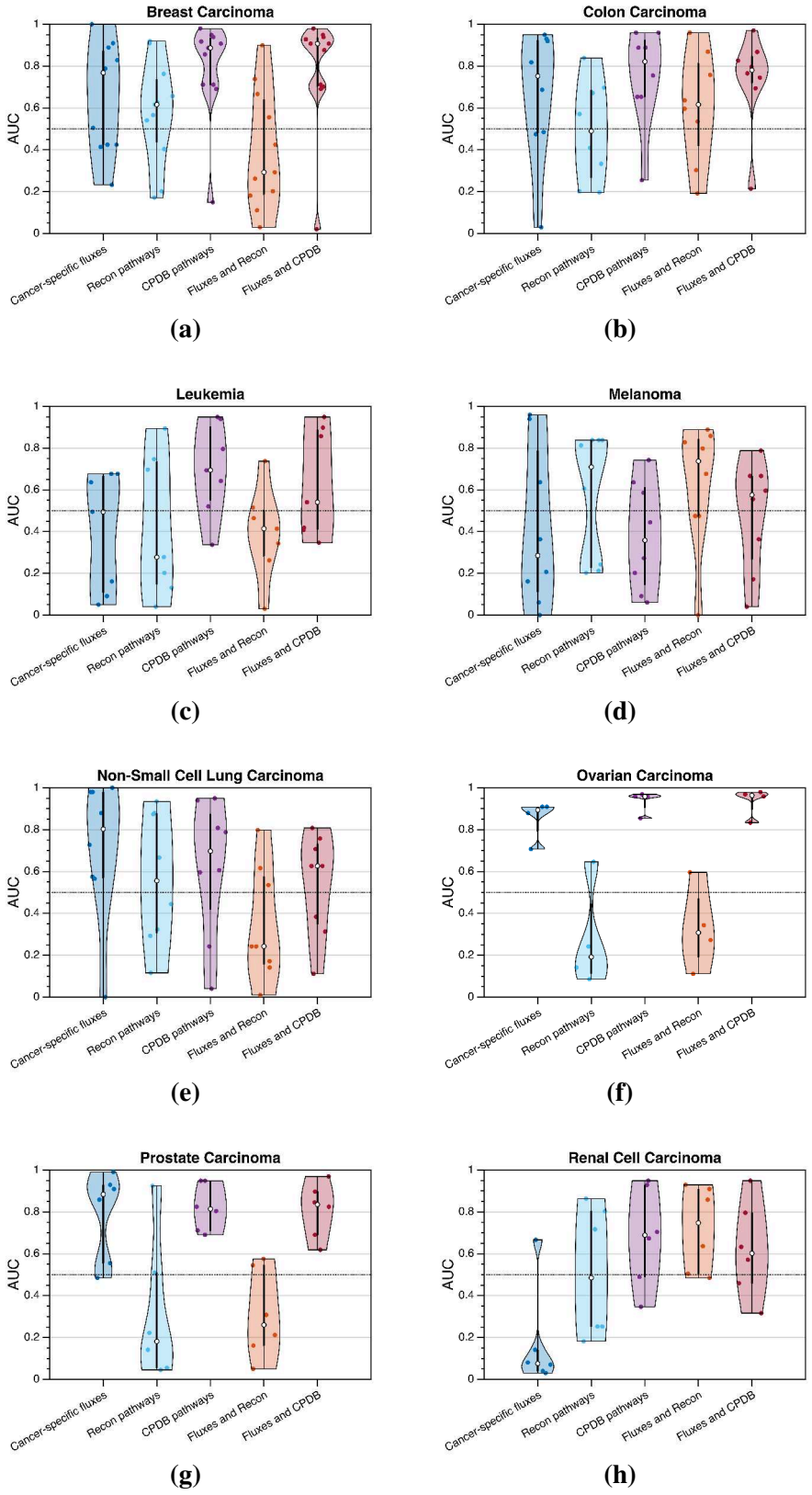


Figure 5.10 | Comparison between flux-based and pathway annotation-based prioritisation - Oncogenes. Comparison of oncogenes AUC distributions among single data sources and two alternative integrations approaches: MKL and CBM. In this case, we consider metabolic flux and pathway annotation data.

Chapter 6

Discussion and future perspectives

6.1 Contributions

Gene prioritisation is progressively becoming essential in molecular biology studies. In fact, we are assisting to a continuous proliferation of a variety of *omic* data brought by technological advances. In the near future it is then likely that more heterogeneous knowledge will have to be combined. Moreover, the classes of biological agents to be prioritised are going to enlarge. For instance, we are only beginning to understand the complex regulation machinery involving non-coding RNA and epigenetic agents. It is estimated that around 90.000 human long non coding genes exist, whose functional implications are progressively emerging [195]. Facing these challenges, the development of novel methods is still strongly needed in order to enhance predictive power and efficiency.

Compared to the considered benchmark kernel methods - MKL1class and ProDiGe - Scuba has some important advantages. ProDiGe is one of the first proposed kernel-based Positive-Unlabelled (PU) learning methods for gene prioritisation and implements a learning strategy based on a biased Support Vector Machine (SVM), which over-weights positive examples during training [95]. In order to reach scalability to large datasets, it leverages a bagging procedure. Like ProDiGe, Scuba implements a learning strategy based on a binary classification set up, but from a different perspective. In a PU problem, the information on positive labels is assumed secure, while the information on negative labels is not. In terms of margin optimisation, this translates in unbalanced entropy on the probability distributions associated to the two sets of training examples. It is then required to regularise more on the unlabelled class - having higher entropy - and in the limit of maximum uncertainty we get the uniform distribution.

MKL1class implements another effective approach for data integration, namely single class learning. This means that the model is obtained solely based on the distribution

of known disease genes, disregarding unlabelled ones. Scuba has enhanced scalability compared to MKL1class, as it involves the optimisation of the 1-norm of the margin vector from the different kernels. In contrast, MKL1class optimises its 2-norm, which is more computationally demanding. Importantly, another distinctive feature of Scuba is a time complexity dependent on the number of positive examples and not on the number of total examples. As a consequence, Scuba can exploit the information on the whole data distribution and at the same time scale to large datasets without the need of sub-sampling the examples. This may be of great advantage as typically disease genes are orders of magnitude less numerous than the candidates.

Results from two different evaluation settings show that our proposed method Scuba outperforms many existing methods, particularly in genome-wide analyses. Compared to the two considered existing kernel-based methods, Scuba performances are always higher, and often significantly higher. Moreover, Scuba has two main levels of scalability that make it particularly suitable for gene prioritisation:

- **Scalability on number of kernels**: Scuba is able to deal with a large number of kernels defined on different data sources. As a consequence, it can be useful to get a more unified view of the problem and to build more powerful predicting models.

- **Scalability on number of training examples**: In typical gene prioritisation problems, the number of known disease genes is much smaller than the number of candidates. Scuba is designed to efficiently deal with unbalanced settings and at the same time take advantage of the whole candidates distribution.

Altogether, our results show that Scuba is a valuable tool to achieve efficient prioritisations, especially in large-scale investigations. A detailed overview on the validation results for single diseases is available in Tables 3.1, A.2 and A.3.

As it is visible in Table A.1, performance with multiple kernels might be close to those with single kernels. Nevertheless, feeding multiple kernels into Scuba alleviates the issue of choosing appropriate kernels for each data source, as implemented in our work. Importantly, this strategy can also provide multiple views on the same data and possibly increase performance. Nevertheless caution must be paid since the more kernels are combined and the more parameters have to be learned, thus increasing the risk of over-fitting. We advice then to moderate the number of kernel matrices generated from each data source.

From an omics data point of view, experimental profiles and annotations can be unable to capture the roots of complex pathological states in some cases. Constraint-Based Modelling (CBM) can help combine and contextualise this information, providing a trade-off between modelling precision and computational cost. Here, we investigated this hypothesis by

developing a pipeline to extract gene-specific information out of a GSMM with the goal of prioritising cancer-associated genes.

The pipeline here introduced (Section 4.4) can in principle be modified or expanded in several points. For instance, Metabolic Expectation Propagation (MEP) is currently the state-of-the-art as regards flux space characterisation, but it can be replaced by a novel techniques in the future [165]. Moreover, strong assumptions on simulating genetic perturbations may limit informativeness of obtained data and advances in this context could enhance also prioritisation accuracy. Despite these limitations, our results show that metabolic fluxes can provide complementary information as compared to gene expression and pathway annotations. However, we observe that they are able to better prioritise genes in only a minority of cases. We hypothesise that further model fine-tuning - for instance taking into account uptake and secretion constraints - could reflect also on prioritisation improvement.

6.2 Open questions

The present study solves some questions but creates new ones at the same. In the following, we delineate the main future research directions.

Improving data integration within multiple kernel learning As testified by experimental results, integration of heterogeneous data is not effectively handled by Multiple Kernel Learning (MKL) in some cases. This is observed as averaging of prediction performance of the different data sources. The reason can be that, in MKL, information from the different data sources interacts too weakly. It is necessary to find new strategies to more strongly combine it, for instance via the definition of kernel functions defined over multiple data sources.

Improvement of condition-specific metabolic models In this study, we employed transcriptomic profiles to generate cancer-specific GSMMs, as this data type is usually preferred in prioritisation analyses. However, protein expression could be in principle even more effective to obtain precise models. Moreover, condition-specific GSMMs could be enhanced by integrating information on metabolite uptake and secretion.

Integration of metabolic, signalling and regulatory networks One of the main limitations in using GSMM is that prioritisation is restricted to genes directly involved in metabolism. Although context-specific models take into account genetic regulation, they do not allow us to simulate the knock-out or de-regulation of regulatory genes. Methods that integrate GSMMs with regulatory networks are promising to this end. In particular,

a recently proposed method permits creating an integrated model starting from transcriptional profiles and the topology of the regulatory network [155]. De-regulation strategies could be straightforwardly implemented into this framework.

Understanding the reasons behind predictions Interpretation of predictions is a widespread issue in machine learning, as most generated models are complex and provide no explanation for their predictions. Methods for prediction interpretation exist but most of them entail a drop in prediction accuracy. Only recently, rule mining was used applied to binary classification and it successfully identified data relationships without accuracy loss [196].

Even in candidate gene prioritisation, understanding why a particular gene has been ranked at the top or at the bottom can be important. As an example, it can help generate new hypotheses and plan following experiments, further accelerating research procedures. MKL provides information in terms of weights assigned to kernels, indicating the level of contribution of each dataset to the final prediction. CBM also can have advantages for explaining predictions. Indeed, a model provides mechanistic explanations at a single reaction level. Application of automatic explanation methods to data generated by CBM can therefore provide more detailed information. In the present study, this issue was not investigated, but it will be priority for future research.

Conclusions

In this work, we advanced two methodological novelties for the prioritisation of candidate disease genes based on the integration of multi-omics data.

In a first stage, we proposed a new computational kernel-based method to guide the identification of novel disease genes, called Scuba. Our method takes advantage of complementary biological knowledge by combining heterogeneous data sources. Every source can be transformed by appropriate kernel functions in order to take full advantage of its information. Our original algorithm is scalable relatively to the size of input data, number of kernel transformations and number of training examples. Experimental results support the thesis that it is effective across a large spectrum of disorders and that can be used to prioritise even the whole genome. Scuba only requires a collection of input genes and optionally a set of candidate genes. The simple requirements make it applicable to a wide range of laboratory investigations. Furthermore, it can be potentially employed also in other prioritisation problems, as long as a PU approach and the integration of heterogeneous biological knowledge are needed. A scientific paper presenting Scuba has been submitted and has very recently been accepted for publication [197]. A Python implementation is also freely available at https://github.com/gzampieri/Scuba.

In a second stage, we introduced a novel approach for gene prioritisation that combines for the first time experimental omics information and CBM within a comprehensive data-driven prediction framework. In other words, we applied Scuba to perform gene prioritisation based on the information reconstructed by integrating transcriptomic profiles and a GSMM. Validation on oncogenes and tumour suppressor genes prioritisation demonstrate a striking complementarity between predictions achieved by our method and by Scuba when fed with conventional gene expression or pathway annotation data. However, such complementarity not always translates into improvements in prioritisation accuracy. These findings demonstrate the potential informativeness of CBM for the identification of novel gene-disease associations, but also highlight that further investigations are necessary in order to achieve optimal data integration in every setting. A major limitation of the method is its limited

applicability to the metabolic domain, however it can be potentially adapted to incorporate emerging methods for the modelling of regulatory networks.

Altogether, these results bring novel methodological strategies to determine the genetic basis of human disease.

Appendix A

Additional results of Chapter 3

Kernel type	kernel hyper-parameter	rank median	rank st.dev.	TPR at top 5% (%)	TPR at top 10% (%)	TPR at top 30% (%)	AUC
Genome-wide prioritisations							
$\mathbf{K}_{MD}$	2	11.13	24.36	31.0	47.6	73.8	0.78
	4	11.11	24.80	35.7	45.2	69.0	0.78
	6	12.34	25.16	33.3	45.2	73.8	0.78
	2,4,6	11.02	24.48	33.3	47.6	73.8	0.78
$\mathbf{K}_{RL}$	1	13.41	20.92	28.6	42.9	76.2	0.80
	10	11.77	21.15	31.0	42.9	78.6	0.81
	100	11.67	21.79	28.6	45.2	76.2	0.80
	1,10,100	12.40	20.90	28.6	40.5	76.2	0.80
Candidate set-based prioritisations							
$\mathbf{K}_{MD}$	2	13.20	26.70	23.8	45.2	69.0	0.77
	4	13.20	27.21	23.8	47.6	76.2	0.77
	6	14.19	27.31	23.8	42.9	73.8	0.77
	2,4,6	13.73	26.88	26.2	47.6	73.8	0.77
$\mathbf{K}_{RL}$	1	11.52	22.98	23.8	42.9	73.8	0.79
	10	11.45	23.25	26.2	47.6	76.2	0.80
	100	11.11	23.78	26.2	45.2	76.2	0.79
	1,10,100	11.60	23.15	23.8	40.5	73.8	0.79

Table A.1 | Test genes rank distribution statistics for different kernel combinations in the time-stamp validation. Scuba results in the experimental setting of Börnigen *et al* [119], using String v8.2 as data source and for different choices of kernels. From the third column to the last one: rank median and standard deviation, TPR in the upper 5/10/30% of the ranking and average AUC. Ranks are normalised in order to lie in the interval $]0, 100]$.

Disease	Associated genes	genome-wide AUC	candidate set AUC
Abdominal aortic aneurysm	ENSG00000136848	0.77	0.84
Alcohol dependence	ENSG00000148680	0.98	0.98
Arthrogryposis	ENSG00000152818	0.98	1.0
Asthma	ENSG00000182578	0.93	0.94
Autosomal recessive primary microcephaly	ENSG00000075702	0.41	0.44
Behcet's disease	ENSG00000136634	0.98	0.97
Bipolar schizoaffective disorder	ENSG00000146276 ENSG00000139618	0.97	0.98
Complex heart defect	ENSG00000121068	0.98	1.0
Congenital anomalies of the kidney and the urinary tract	ENSG00000164736 ENSG00000178188	0.97	0.96
Congenital diaphragmatic hernia	ENSG00000004961 ENSG00000154309	0.86	0.87
Crohn's disease	ENSG00000176920 ENSG00000185651 ENSG00000069399	0.89	0.89
Dursun syndrome	ENSG00000141349	0.58	0.46
Ehlers-Danlos syndrome	ENSG00000169105	0.99	1.0
Esophageal squamous cell carcinoma	ENSG00000138193 ENSG00000101276	0.3	0.23
Leprosy	ENSG00000111537	0.96	0.9
Lung adenocarcinoma	ENSG00000073282	0.89	0.84
Methylmalonic aciduria	ENSG00000167775	0.9	0.93
Metopic craniosynostosis	ENSG00000106571	0.98	0.98
Mitochondrial complex I deficiency	ENSG00000177646	0.95	0.96
Multiple sclerosis	ENSG00000120088	0.83	0.84
Myelodysplastic syndromes	ENSG00000106462	0.81	0.83
Nasopharyngeal carcinoma	ENSG00000085276 ENSG00000127863	0.81	0.8
Nonsyndromic cleft lip/palate	ENSG00000148175	0.82	0.8
Parkinson's disease	ENSG00000175104	0.82	0.8
Periventricular heterotopia	ENSG00000102103	0.54	0.45
Primary biliary cirrhosis	ENSG00000142606 ENSG00000142539	0.82	0.77
Psoriasis	ENSG00000056972	0.96	1.0
Retinal-renal ciliopathy	ENSG00000054282	1.0	1.0
Single-suture craniosynostosis	ENSG00000124813	0.98	0.98
Smooth pursuit eye movement abnormality	ENSG00000099901	0.27	0.2
Testicular germ cell tumor	ENSG00000137090 ENSG00000171681	0.5	0.41
Tetralogy of Fallot	ENSG00000145012	0.74	0.67
Type 2 diabetes	ENSG00000182247	0.21	0.19

Table A.2 | Test genes AUC for individual disorders in the time-stamp validation. Scuba performance for single disorders considered by Börnigen *et al* in their evaluation of gene prioritisation tools [119].

Disease	Associated genes	genome-wide AUC
Behcet's disease	ENSG00000162594 ENSG00000168811 ENSG00000138378 ENSG00000163823 ENSG00000136869 ENSG00000183542 ENSG00000164307 ENSG00000026103 ENSG00000206340 ENSG00000206450 ENSG00000134882	0.87
Bipolar schizoaffective disorder	ENSG00000175344 ENSG00000138592 ENSG00000151702 ENSG00000124782 ENSG00000171988 ENSG00000176986	0.68
Crohn's disease	ENSG00000140368	0.90
Parkinson's disease	ENSG00000064692 ENSG00000153234 ENSG00000116675 ENSG00000159082 ENSG00000184381 ENSG00000138246	0.89
Primary biliary cirrhosis	ENSG00000128604 ENSG00000181634 ENSG00000105329 ENSG00000110777 ENSG00000064419 ENSG00000016602 ENSG00000141076 ENSG00000106089 ENSG00000132912	0.76
Psoriasis	ENSG00000206237 ENSG00000196126 ENSG00000179344 ENSG00000206306 ENSG00000163599 ENSG00000206240 ENSG00000077150 ENSG00000141527 ENSG00000198246	0.94
Smooth pursuit eye movement abnormality	ENSG00000104133 ENSG00000171385 ENSG00000020922 ENSG00000070610 ENSG00000013503 ENSG00000167658	0.73

Table A.3 | Test genes AUC for individual multi-factorial disorders in the expanded time-stamp validation. Scuba performance for single disorders considered in Table 3.5 in the main text. These are the multi-factorial diseases employed by Börnigen *et al* [119] with at least a new gene annotation between March 2013 and February 2017 as reported by the Human Phenotype Ontology [71].

Appendix B

Additional results of Chapter 5

Table B.1 | Full list of metabolic reactions whose predicted fluxes correlate with measured proliferation. Recon reactions significantly correlated to tumour cells proliferation obtained from the NCI60 panel [169].

Pathway	Reaction	Reaction name	Pearson r	p-value
	r1411_f	D-Galactosyl-N-acetyl-D-galactosaminyl-(N-acetylneuraminyl)-D-galactosyl-D-glucosylceramide galactohydrolase EC:3.2.1.23	-0.47	2.94E-04
	EX_fum(e)_f	Exchange of Fumarate	0.39	2.94E-03
	biomass_reaction	Generic human biomass reaction	0.41	1.56E-03
	biomass_protein	protein component of biomass	0.41	1.56E-03
	biomass_DNA	DNA component of biomass	0.41	1.56E-03
	biomass_RNA	RNA component of biomass	0.41	1.56E-03
	biomass_carbohydrate	carbohydrate component of biomass	0.41	1.56E-03
	biomass_lipid	lipid component of biomass	0.41	1.56E-03
	biomass_other	other component of biomass	0.41	1.56E-03
	BETBGTtc_f	betaine transport by BGT	0.40	2.06E-03
	FOLOATPtc_f	folate transport by OATP	0.40	2.34E-03
	GLNB0AT3tc	glutamine transport by B0AT3	0.36	6.86E-03
	H2OGLYAQPt_f	water and glycerol transport by AQP	0.50	8.35E-05
	INSt2_f	inosine transport in via proton symport, reversible	0.37	4.58E-03
	OCDCAFAPMtc	octadecanoate tranport by FAT	0.38	4.10E-03
	ATVACIDOATPtu_f	uptake of atorvastatin by enterocytes	0.35	9.21E-03
	ATVACIDtdu_f	passive diffusion of atorvastatin into enterocytes	0.39	2.74E-03
	LST4EXPTDhc_f	uptake of Losartan-E3174 into hepatocytes via diffusion	0.38	4.12E-03
	LSTNM7TDhc_f	uptake of Losartan-N2-glucuronide / Losartan-M7 into hepatocytes via diffusion	0.38	4.12E-03
	LSTNtd_f	uptake of losartan via diffusion into enterocytes	0.38	4.12E-03
	OXYPthc_f	uptake of oxypurinol by hepatocytes	0.40	2.52E-03
	OXYPtepv_f	release of oxypurinol into portal blood	0.40	2.52E-03
	PVSOATPtu_f	uptake of pravastatin by enterocytes	0.37	4.65E-03
	TMDOATPtsc_f	uptake of torasemide into enterocytes via antiport	0.36	5.87E-03
	TMDtd_f	uptake of torasemide into enterocytes via diffusion	0.40	2.19E-03
	r0390_b	Isomaltose 6-alpha-D-glucanohydrolase Starch and sucrose metabolism EC:3.2.1.10	-0.35	8.44E-03
	r0737_b	3-Ketolactose galactohydrolase Galactose metabolism EC:3.2.1.23	0.35	7.87E-03
	fumt_b	Fumarate transport	0.39	2.94E-03
	CLCFTRte_b	chloride transport by CFTR	0.38	4.12E-03
	DOPAENT4tc_b	dopa transport by ENT4	0.36	6.99E-03
	SRTNENT4tc_b	serotonin transport by ENT4	0.40	2.53E-03

Pathway	Reaction	Reaction name	Pearson *r*	p-value
	ATVACIDMCTtu_b	uptake of atorvastatin by enterocytes by MCT1	0.37	4.80E-03
	LST4EXPthc_b	uptake of Losartan-E3174 into hepatocytes	0.38	4.12E-03
	LSTNM7thc_b	uptake of Losartan-N2-glucuronide / Losartan-M7 into hepatocytes	0.38	4.12E-03
	LSTNRATt_b	uptake of losartan via antiport into enterocytes	0.38	4.12E-03
	PVSHtu_b	uptake of pravastatin by enterocytes via proton coupled mechanism	0.39	3.14E-03
	PVStep_b	pravastatin exit into portal blood	0.41	1.91E-03
	TMDOAThc_b	uptake of torasemide into hepatocytes via OAT1	0.37	4.84E-03
Aminosugar metabolism	CHTNASEe	chitinase, extracellular	-0.40	2.55E-03
Aminosugar metabolism	r1374	EC:3.2.1.14	-0.40	2.55E-03
Butanoate metabolism	BDHm_b	(R)-3-Hydroxybutanoate:NAD+ oxidoreductase	0.36	6.40E-03
Cholesterol metabolism	C14STRr	C-14 sterol reductase	0.41	1.56E-03
Cholesterol metabolism	C3STDH1Pr	C-3 sterol dehydrogenase (4-methylzymosterol)	0.41	1.56E-03
Cholesterol metabolism	C4STMO1r	C-4 sterol methyl oxidase (4,4-dimethylzymosterol)	0.41	1.56E-03
Cholesterol metabolism	DMATT	dimethylallyltranstransferase	0.39	2.92E-03
Cholesterol metabolism	EBP1r	3-beta-hydroxysteroid-delta(8),delta(7)-isomerase	0.37	5.67E-03
Cholesterol metabolism	EBP2r	3-beta-hydroxysteroid-delta(8),delta(7)-isomerase	0.36	6.49E-03
Cholesterol metabolism	GRTT	geranyltranstransferase	0.39	2.92E-03
Cholesterol metabolism	LNSTLSr	lanosterol synthase	0.41	1.56E-03
Cholesterol metabolism	SQLEr	Squalene epoxidase, endoplasmic reticular (NADP)	0.41	1.56E-03
Cholesterol metabolism	SQLSr	Squalene synthase	0.41	1.56E-03
Cholesterol metabolism	r0170_f	Farnesyl-diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21	0.41	1.56E-03
Cholesterol metabolism	r0575_f	Presqualene diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21	0.41	1.56E-03
Cholesterol metabolism	r0781_f	Lanosterol,NADPH:oxygen oxidoreductase (14-methyl cleaving) Biosynthesis of steroids EC:1.14.13.70	0.41	1.56E-03
Citric acid cycle	r0081_f	L-Alanine:2-oxoglutarate aminotransferase Glutamate metabolism / Alanine and aspartate metabolism EC:2.6.1.2	0.46	3.49E-04
Exchange/demand reaction	DM_datp_m_	dATP demand	0.35	9.17E-03
Exchange/demand reaction	DM_btn	Demand for biotin	-0.36	6.67E-03
Exchange/demand reaction	EX_5mthf(e)_f	exchange reaction for 5-Methyltetrahydrofolate	0.42	1.26E-03
Exchange/demand reaction	EX_ac(e)_f	Acetate exchange	0.38	3.63E-03
Exchange/demand reaction	EX_acgam(e)_f	N-Acetyl-D-glucosamine exchange	-0.40	2.55E-03
Exchange/demand reaction	EX_adprbp(e)_f	ADPribose 2-phosphate exchange	-0.37	4.52E-03
Exchange/demand reaction	EX_ak2lgchol_hs(e)_f	1-alkyl 2-lysoglycerol 3-phosphocholine exchange	-0.38	4.20E-03
Exchange/demand reaction	EX_bhb(e)_f	(R)-3-Hydroxybutanoate transport via H+ symport	0.36	6.40E-03

Pathway	Reaction	Reaction name	Pearson *r*	p-value
Exchange/demand reaction	EX_cl(e)_f	exchange reaction for Chloride	-0.35	8.31E-03
Exchange/demand reaction	EX_co2(e)_f	CO2 exchange	0.35	7.56E-03
Exchange/demand reaction	EX_elaid(e)_f	elaidic acid exchange	0.39	3.21E-03
Exchange/demand reaction	EX_glc(e)_f	D-Glucose exchange	-0.39	2.92E-03
Exchange/demand reaction	EX_glyc(e)_f	Glycerol exchange	-0.37	5.44E-03
Exchange/demand reaction	EX_h2o(e)_f	H2O exchange	-0.40	2.55E-03
Exchange/demand reaction	EX_leuktrD4(e)_f	leukotriene D4 exchange	0.37	4.90E-03
Exchange/demand reaction	EX_malt(e)_f	Maltose exchange	-0.39	2.80E-03
Exchange/demand reaction	EX_ncam(e)_f	Nicotinamide exchange	-0.38	4.17E-03
Exchange/demand reaction	EX_Tyr_ggn(e)_f	Tyr-194 of apo-glycogenin protein (primer for glycogen synthesis) exchange	-0.39	2.80E-03
Exchange/demand reaction	EX_uri(e)_f	exchange reaction for Uridine	0.37	5.17E-03
Exchange/demand reaction	EX_HC00229(e)_f	Exchange of Isomaltose	-0.35	8.44E-03
Exchange/demand reaction	EX_HC01446(e)_f	Exchange of 3-Ketolactose	0.35	7.87E-03
Exchange/demand reaction	EX_HC02160(e)_f	Exchange of GM2-pool	-0.47	2.94E-04
Exchange/demand reaction	EX_no2(e)_f	Nitrite exchange	-0.38	3.46E-03
Exchange/demand reaction	EX_HC00822(e)_f	Chitobiose exchange	-0.40	2.55E-03
Exchange/demand reaction	EX_3ump(e)_f	3-UMP(2-) exchange	-0.38	4.24E-03
Exchange/demand reaction	EX_3octdeccrn__f	exchange reaction for 3-hydroxyoctadecanoyl carnitine	0.37	4.49E-03
Exchange/demand reaction	EX_cysam(e)_f	exchange reaction for cysam	-0.41	1.89E-03
Exchange/demand reaction	EX_glyc3p(e)_f	exhange of glycerol 3-phosphate	0.36	5.76E-03
Exchange/demand reaction	EX_acac(e)_b	Acetoacetate exchange	0.45	4.79E-04
Exchange/demand reaction	EX_chtn(e)_b	chitin exchange	-0.40	2.55E-03
Exchange/demand reaction	EX_glu_L(e)_b	L-Glutamate exchange	0.45	5.87E-04
Exchange/demand reaction	EX_glygn4(e)_b	exchange reaction for glycogen, structure 4 (glycogenin-1,6-{2[1,4-Glc], [1,4-Glc]})	-0.43	8.39E-04
Exchange/demand reaction	EX_glygn5(e)_b	exchange reaction for glycogen, structure 5 (glycogenin-2[1,4-Glc])	-0.40	2.07E-03
Exchange/demand reaction	EX_h2o2(e)_b	Hydrogen peroxide exchange	0.46	3.70E-04
Exchange/demand reaction	EX_his_L(e)_b	exchange reaction for L-histidine	0.41	1.56E-03
Exchange/demand reaction	EX_ins(e)_b	Inosine exchange	0.35	7.50E-03
Exchange/demand reaction	EX_leuktrF4(e)_b	leukotriene F4 exchange	0.36	6.07E-03
Exchange/demand reaction	EX_lys_L(e)_b	L-Lysine exchange	0.55	1.13E-05
Exchange/demand reaction	EX_nad(e)_b	Nicotinamide adenine dinucleotide exchange	-0.37	4.54E-03
Exchange/demand reaction	EX_nadp(e)_b	Nicotinamide adenine dinucleotide phosphate exchange	-0.37	4.52E-03
Exchange/demand reaction	EX_paf_hs(e)_b	1-alkyl 2-acteylglycerol 3-phosphocholine (homo sapiens) exchange	-0.38	4.20E-03
Exchange/demand reaction	EX_pglyc_hs(e)_b	phosphatidylglycerol (homo sapiens) exchange	0.41	1.56E-03
Exchange/demand reaction	EX_thr_L(e)_b	L-Threonine exchange	0.41	1.56E-03
Exchange/demand reaction	EX_trp_L(e)_b	L-Tryptophan exchange	0.41	1.56E-03

Pathway	Reaction	Reaction name	Pearson *r*	p-value
Exchange/demand reaction	EX_HC01440(e)_b	Exchange of 3-Keto-beta-D-galactose	0.35	7.87E-03
Exchange/demand reaction	EX_HC02161(e)_b	Exchange of GM1-pool	-0.47	2.94E-04
Exchange/demand reaction	EX_CE2011(e)_b	hypothiocyanite exchange	-0.38	4.37E-03
Exchange/demand reaction	EX_23cump(e)_b	2,3-cyclic UMP(1-) exchange	-0.38	4.24E-03
Exchange/demand reaction	EX_CE4881(e)_b	nitryl chloride exchange	-0.38	3.46E-03
Exchange/demand reaction	EX_CE0074(e)_b	alloxan exchange	0.40	2.53E-03
Exchange/demand reaction	EX_ptth(e)_b	exchange reaction for ptth	-0.41	1.89E-03
Fatty acid oxidation	C180CPT1	carnitine acyltransferase I	0.37	4.49E-03
Fatty acid oxidation	C180CPT2	carnitine acyltransferase II	0.37	4.49E-03
Fatty acid oxidation	C18OHc_f	production of 3-hydroxyoctadecanoylcarnitine	0.37	4.49E-03
Fatty acid oxidation	FAOXC18C18OHm	fatty acid beta oxidation(C18→ C18OH)m	0.37	4.49E-03
Fatty acid oxidation	C180CRNt	carnitine/acylcarnitine translocase	0.37	4.49E-03
Fatty acid oxidation	HOCDACBP_f	transport of (S)-3-Hydroxyoctadecanoyl-CoA from mitochondria into the cytosol	0.37	4.49E-03
Fatty acid oxidation	HOCTDECCRNe	transport of 3-hydroxyoctadecanoyl carnitine into extra cellular space	0.37	4.49E-03
Fatty acid oxidation	FACOAL1813_b	fatty-acid–CoA ligase	0.39	3.21E-03
Fatty acid synthesis	DESAT18_5	stearoyl-CoA desaturase (n-C18:0CoA ->n-C18:1CoA)	0.40	2.43E-03
Folate metabolism	MTHFR3	5,10-methylenetetrahydrofolatereductase (NADPH)	0.39	2.92E-03
Folate metabolism	MTHFD2m_b	methylenetetrahydrofolate dehydrogenase (NAD), mitochondrial	0.40	2.39E-03
Folate metabolism	r0792_b	5-methyltetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate / Methane metabolism EC:1.5.1.20	0.40	2.26E-03
Glutamate metabolism	GLUDym_f	glutamate dehydrogenase (NADP), mitochondrial	0.35	7.66E-03
Glycerophospholipid metabolism	CLS_hs	cardiolipin synthase (homo sapiens)	0.41	1.56E-03
Glycerophospholipid metabolism	PAFHe	Platelet-activating factor acetylhydrolase	-0.36	5.86E-03
Glycerophospholipid metabolism	G3PD1_b	glycerol-3-phosphate dehydrogenase (NAD)	0.35	7.68E-03
Glycine, serine, alanine and threonine metabolism	GHMT2r_f	glycine hydroxymethyltransferase, reversible	0.34	9.46E-03
Miscellaneous	RE0702E	RE0702	-0.38	4.37E-03
Miscellaneous	RE1860E_f	RE1860	-0.38	4.24E-03
Miscellaneous	RE2513E	RE2513	-0.35	8.31E-03
Miscellaneous	RE2514E_b	RE2514	-0.38	3.46E-03
NAD metabolism	NADNe	NAD nucleosidase,extracellular	-0.37	4.54E-03
NAD metabolism	NADPNe	NADP nucleosidase,extracellular	-0.37	4.52E-03
Purine catabolism	PUNP5_f	purine-nucleoside phosphorylase (Inosine)	0.35	9.07E-03
Purine catabolism	RE2888E	RE2888	0.40	2.53E-03
Pyrimidine catabolism	D3AIBTm_b	D-3-Amino-isobutanoate:pyruvate aminotransferase, mitochondrial	0.34	9.68E-03
Sphingolipid metabolism	DHCRD1	dihydroceramide desaturase	0.36	6.23E-03

Pathway	Reaction	Reaction name	Pearson *r*	p-value
Sphingolipid metabolism	DSAT	dihydrosphingosine N-acyltransferase	0.41	1.56E-03
Sphingolipid metabolism	SMS	Sphingomyelin synthase (homo sapiens)	0.41	1.56E-03
Squalene and cholesterol synthesis	HMGCOARc	Hydroxymethylglutaryl CoA reductase (ir) in cytosol	0.41	1.56E-03
Squalene and cholesterol synthesis	IPDDI_f	isopentenyl-diphosphate D-isomerase	0.39	2.92E-03
Starch and sucrose metabolism	GAMYe	glucoamylase, extracellular (glygn5 ->malt)	-0.39	2.80E-03
Transport, endoplasmic reticular	CHSTEROLtrc_f	transport of cholesterol into the cytosol	0.41	1.56E-03
Transport, endoplasmic reticular	FRDPtcr_f	transport of Farnesyl diphosphate into the endoplasmic reticulum	0.42	1.28E-03
Transport, endoplasmic reticular	FORtr_b	FOR transporter, endoplasmic reticulum	0.41	1.56E-03
Transport, endoplasmic reticular	r1051_b	Vesicular transport	0.41	1.56E-03
Transport, extracellular	PNTEHe	PNTEHe	-0.41	1.89E-03
Transport, extracellular	ADNt_f	adenosine facilated transport in cytosol	0.37	4.81E-03
Transport, extracellular	ADNt4	adenosine transport (Na/Adn cotransport)	0.40	2.08E-03
Transport, extracellular	ARACHDt2_f	fatty acid transport via diffusion	0.38	3.92E-03
Transport, extracellular	ARACHt	arachidate transport by FAT	0.40	2.05E-03
Transport, extracellular	ASCBt_f	L-ascorbate transport via facilitated diffusion	0.38	4.34E-03
Transport, extracellular	CLHCO3tex2	chloride transport via bicarbonate countertransport (2:1)	0.35	8.22E-03
Transport, extracellular	DOPAtu_f	Dopamine uniport	0.40	2.14E-03
Transport, extracellular	FATP3t_f	fatty acid electroneutral transport	0.40	2.49E-03
Transport, extracellular	FATP4t_f	fatty acid electroneutral transport	0.40	2.49E-03
Transport, extracellular	FATP5t_f	fatty acid electroneutral transport	0.39	3.09E-03
Transport, extracellular	FATP6t_f	fatty acid electroneutral transport	0.36	5.90E-03
Transport, extracellular	FATP9t_f	fatty acid electroneutral transport	0.41	1.53E-03
Transport, extracellular	INSt_f	Inosine transport (diffusion)	0.42	1.40E-03
Transport, extracellular	LYStiDF	L-lysine transport via diffusion (extracellular to cytosol)	0.34	9.24E-03
Transport, extracellular	METtec_f	L-methionine transport via diffusion (extracellular to cytosol)	0.34	9.71E-03
Transport, extracellular	NRPPHRtu_f	Norepinephrine uniport	0.39	2.73E-03
Transport, extracellular	PGLYCt_f	phosphatidylglycerol transport	0.41	1.56E-03
Transport, extracellular	PHEtec_f	L-phenylalanine transport via diffusion (extracellular to cytosol)	0.35	8.57E-03
Transport, extracellular	SRTNtu_f	Serotonin uniport	0.39	2.60E-03
Transport, extracellular	THYMDt1	thymd transport	0.37	4.74E-03
Transport, extracellular	TRPt_f	L-tryptophan transport	0.34	9.52E-03
Transport, extracellular	r0839_f	Postulated transport reaction	0.43	9.26E-04
Transport, extracellular	r1144	Major Facilitator(MFS) TCDB:2.A.18.6.3	0.37	4.86E-03
Transport, extracellular	r1551_f	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.34	9.89E-03
Transport, extracellular	r1557_f	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.34	9.79E-03
Transport, extracellular	r1571_f	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.36	6.21E-03

Pathway	Reaction	Reaction name	Pearson r	p-value
Transport, extracellular	r1584_f	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.35	8.06E-03
Transport, extracellular	r1625_f	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.38	3.88E-03
Transport, extracellular	r1659_f	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.37	5.03E-03
Transport, extracellular	r1660_f	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.42	1.19E-03
Transport, extracellular	r1665_f	Y+LAT2 Utilized transport	0.41	1.52E-03
Transport, extracellular	r1666_f	Y+LAT2 Utilized transport	0.42	1.44E-03
Transport, extracellular	r2101_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.44	6.49E-04
Transport, extracellular	r2102_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.37	4.46E-03
Transport, extracellular	r2105_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.40	2.13E-03
Transport, extracellular	r2106_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.35	9.04E-03
Transport, extracellular	r2107_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.39	2.64E-03
Transport, extracellular	r2108_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.39	3.04E-03
Transport, extracellular	r2120_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.35	7.69E-03
Transport, extracellular	r2125_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.44	6.59E-04
Transport, extracellular	r2126_f	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.43	8.25E-04
Transport, extracellular	r2218_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.39	3.00E-03
Transport, extracellular	r2219_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.39	3.00E-03
Transport, extracellular	r2220_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.39	3.00E-03
Transport, extracellular	r2251_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.41	1.71E-03
Transport, extracellular	r2252_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.41	1.71E-03
Transport, extracellular	r2253_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.41	1.71E-03
Transport, extracellular	r2284_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.35	8.44E-03
Transport, extracellular	r2285_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.35	8.44E-03
Transport, extracellular	r2286_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.35	8.44E-03
Transport, extracellular	r2309_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	0.37	5.37E-03
Transport, extracellular	r2312_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	0.42	1.43E-03
Transport, extracellular	r2525_f	Major Facilitator(MFS) TCDB:2.A.1.44.1	0.37	5.00E-03
Transport, extracellular	r2532_f	Major Facilitator(MFS) TCDB:2.A.1.44.1	0.35	7.50E-03
Transport, extracellular	r2534_f	Major Facilitator(MFS) TCDB:2.A.1.44.1	0.36	6.81E-03
Transport, extracellular	r2535_f	Major Facilitator(MFS) TCDB:2.A.1.44.1	0.35	9.08E-03
Transport, extracellular	CYSSNAT5tc_f	transport of L-Cysteine into the cell coupled with co-transport with Sodium and counter transport with proton by SNAT5 transporter.	0.36	6.31E-03
Transport, extracellular	GLYSNAT5tc_f	transport of Glycine into the cell coupled with co-transport with Sodium and counter transport with proton by SNAT5 transporter.	0.36	7.06E-03
Transport, extracellular	5MTHFt2_b	5-methyltetrahydrofolate transport via anion exchange	0.39	3.11E-03
Transport, extracellular	BTNt2_b	Biotin reversible transport via proton symport	0.38	3.61E-03

Pathway	Reaction	Reaction name	Pearson *r*	p-value
Transport, extracellular	CO2t_b	CO2 transporter via diffusion	0.43	1.02E-03
Transport, extracellular	DOPAt4_2_r_b	Dopamine reversible transport in via sodium symport (1:2)	0.39	3.08E-03
Transport, extracellular	ELAIDt_b	fatty acid transport via diffusion	0.39	3.21E-03
Transport, extracellular	FOLt2_b	folate transport via anion exchange	0.35	7.67E-03
Transport, extracellular	GLYBt4_2_r_b	Betaine transport (sodium symport) (2:1)	0.43	1.06E-03
Transport, extracellular	MANt4_b	D-mannose transport via sodium cotransport	0.39	2.74E-03
Transport, extracellular	NAIt_b	Na+ / iodide cotransport	0.39	3.00E-03
Transport, extracellular	NAt_b	sodium transport (uniport)	0.40	2.49E-03
Transport, extracellular	NRPPHRt4_2_r_b	Norepinephrine reversible transport in via sodium symport (1:2)	0.39	2.78E-03
Transport, extracellular	PRODt2r_b	D-proline reversible transport via proton symport	0.40	2.52E-03
Transport, extracellular	r0942_b	Neurotransmitter:Sodium Symporter (NSS) TCDB:2.A.22.3.4	0.40	2.51E-03
Transport, extracellular	r1608_b	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.39	3.23E-03
Transport, extracellular	r1615_b	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.41	1.91E-03
Transport, extracellular	r1654_b	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.1	0.37	4.80E-03
Transport, extracellular	r2080_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.48	1.87E-04
Transport, extracellular	r2081_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.44	6.64E-04
Transport, extracellular	r2082_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.51	5.72E-05
Transport, extracellular	r2083_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.40	2.20E-03
Transport, extracellular	r2084_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.45	4.70E-04
Transport, extracellular	r2085_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.45	5.01E-04
Transport, extracellular	r2087_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.41	1.67E-03
Transport, extracellular	r2110_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.41	1.91E-03
Transport, extracellular	r2113_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.41	1.64E-03
Transport, extracellular	r2114_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	0.35	7.56E-03
Transport, extracellular	r2233_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.42	1.36E-03
Transport, extracellular	r2234_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.42	1.36E-03
Transport, extracellular	r2235_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	0.42	1.36E-03
Transport, extracellular	r2485_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	0.39	2.92E-03
Transport, extracellular	r2486_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	0.39	2.92E-03
Transport, extracellular	r2487_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	0.39	2.92E-03
Transport, extracellular	r2494_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	0.44	6.55E-04
Transport, extracellular	r2495_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	0.44	6.55E-04
Transport, extracellular	r2496_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	0.44	6.55E-04
Transport, extracellular	GLYt7_311_r_b	glycine reversible transport via sodium and chloride symport (3:1:1)	0.36	6.66E-03
Transport, extracellular	HCO3_NAt_b	bicarbonate transport (Na/HCO3 cotransport)	0.35	7.97E-03
Transport, extracellular	GLYCTDle_b	difussion of glycerol accross the brush border membrane	0.46	3.58E-04

Pathway	Reaction	Reaction name	Pearson r	p-value
Transport, extracellular	PRO_Dtde_b	D-proline transport, extracellular	0.40	2.52E-03
Transport, extracellular	glyc3pte_b	glycerol 3-phosphate transport	0.36	5.76E-03
Transport, lysosomal	r1052	Vesicular transport	0.41	1.56E-03
Transport, mitochondrial	r1146_f	Biosynthesis of steroids Enzyme catalyzed	0.41	1.56E-03
Transport, mitochondrial	CHSTEROLt2_f	cholesterol intracellular transport	0.41	1.56E-03
Transport, mitochondrial	PROtm_f	L-proline transport, mitochondrial	0.38	3.99E-03
Transport, mitochondrial	ACt2m_b	acetate mitochondrial transport via proton symport	0.38	4.28E-03
Transport, mitochondrial	BHBtm_b	(R)-3-Hydroxybutanoate mitochondrial transport via H+ symport	0.36	6.40E-03
Transport, mitochondrial	CHSTEROLt3_b	cholesterol intracellular transport	0.41	1.56E-03
Transport, mitochondrial	THFtm_b	5,6,7,8-Tetrahydrofolate transport, diffusion, mitochondrial	0.41	1.52E-03
Transport, mitochondrial	r0911_b	Facilitated diffusion	0.35	9.10E-03
Transport, peroxisomal	FRDPtr_f	lipid, flip-flop intracellular transport	0.38	3.93E-03

Table B.2 | Full list of metabolic reactions whose predicted fluxes correlate with breast cancer patient survival. Recon reactions significantly correlated to breast cancer patient survival obtained from the GDAC repository [191].

Pathway	Reaction	Reaction name	Pearson *r*	p-value
	r0267_f	CMP-N-acetylneuraminate,ferrocytochrome-b5:oxygen oxidoreductase (N-acetyl-hydroxylating) Aminosugars metabolism EC:1.14.18.2	-0.27	5.99E-03
	r0407_f	Sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase Carbon fixation EC:4.1.2.13	-0.29	2.44E-03
	r0610_f	CTP:D-Tagatose 6-phosphate 1-phosphotransferase Galactose metabolism EC:2.7.1.11	-0.26	8.13E-03
	r1135_f	hydroxysteroid (17-beta) dehydrogenase 7 Biosynthesis of steroids EC:1.1.1.270	-0.27	5.03E-03
	biomass_reaction	Generic human biomass reaction	-0.27	5.03E-03
	biomass_protein	protein component of biomass	-0.27	5.03E-03
	biomass_DNA	DNA component of biomass	-0.27	5.03E-03
	biomass_RNA	RNA component of biomass	-0.27	5.03E-03
	biomass_carbohydrate	carbohydrate component of biomass	-0.27	5.03E-03
	biomass_lipid	lipid component of biomass	-0.27	5.03E-03
	biomass_other	other component of biomass	-0.27	5.03E-03
	r0268_b	cytidine monophospho-N-acetylneuraminic acid hydroxylase EC:1.14.18.2	-0.27	5.99E-03
	r0611_b	ITP:D-Tagatose 6-phosphate 1-phosphotransferase Galactose metabolism EC:2.7.1.11	-0.26	8.13E-03
Arginine and Proline Metabolism	r0617_f	trans-4-Hydroxy-L-proline:NADP+ 5-oxidoreductase Arginine and proline metabolism EC:1.5.1.2	-0.27	5.99E-03
Arginine and Proline Metabolism	r0615_b	trans-4-Hydroxy-L-proline:NAD+ 5-oxidoreductase Arginine and proline metabolism EC:1.5.1.2	-0.27	5.99E-03
Bile acid synthesis	AKR1C41	aldo-keto reductase family 1, member C4 (chlordecone reductase; 3-alpha hydroxysteroid dehydrogenase, type I; dihydrodiol dehydrogenase 4)	-0.27	5.99E-03
Bile acid synthesis	AKR1C42	aldo-keto reductase family 1, member C4 (chlordecone reductase; 3-alpha hydroxysteroid dehydrogenase, type I; dihydrodiol dehydrogenase 4)	-0.27	5.99E-03
Bile acid synthesis	r0747_f	3alpha,7alpha-Dihydroxy-5beta-cholestane:NADP+ oxidoreductase (B-specific); 3alpha,7alpha-Dihydroxy-5beta-cholestane:NADP+ oxidoreductase Bile acid biosynthesis EC:1.1.1.50	-0.27	5.99E-03
Bile acid synthesis	r0750_f	3alpha,7alpha,12alpha-Trihydroxy-5beta-cholestane:NADP+ oxidoreductase (B-specific); 3alpha,7alpha,12alpha-Trihydroxy-5beta-cholestane:NADP+ oxidoreductase Bile acid biosynthesis EC:1.1.1.50	-0.27	5.99E-03
Bile acid synthesis	RE1807C_f	RE1807	-0.27	5.99E-03
Bile acid synthesis	RE2626C_f	RE2626	-0.27	5.99E-03
Bile acid synthesis	RE3346C_f	RE3346	-0.27	5.99E-03

Pathway	Reaction	Reaction name	Pearson *r*	p-value
Bile acid synthesis	r0688_b	3alpha,7alpha-Dihydroxy-5beta-cholestan-26-al:NAD+ oxidoreductase Bile acid biosynthesis EC:1.2.1.3	-0.27	5.99E-03
Cholesterol metabolism	C14STRr	C-14 sterol reductase	-0.27	5.03E-03
Cholesterol metabolism	C4STMO1r	C-4 sterol methyl oxidase (4,4-dimethylzymosterol)	-0.27	5.03E-03
Cholesterol metabolism	DHCR71r	7-dehydrocholesterol reductase	-0.27	5.20E-03
Cholesterol metabolism	EBP1r	3-beta-hydroxysteroid-delta(8),delta(7)-isomerase	-0.29	3.01E-03
Cholesterol metabolism	LNSTLSr	lanosterol synthase	-0.27	5.03E-03
Cholesterol metabolism	LSTO1r	Lathosterol oxidase	-0.27	5.20E-03
Cholesterol metabolism	SQLEr	Squalene epoxidase, endoplasmic reticular (NADP)	-0.27	5.03E-03
Cholesterol metabolism	SQLSr	Squalene synthase	-0.27	5.03E-03
Cholesterol metabolism	r0170_f	Farnesyl-diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21	-0.27	5.03E-03
Cholesterol metabolism	r0575_f	Presqualene diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21	-0.27	5.03E-03
Cholesterol metabolism	r0781_f	Lanosterol,NADPH:oxygen oxidoreductase (14-methyl cleaving) Biosynthesis of steroids EC:1.14.13.70	-0.27	5.03E-03
Eicosanoid metabolism	RE3566C_b	RE3566	-0.27	5.65E-03
Exchange/demand reaction	EX_for(e)_f	Formate exchange	-0.27	5.03E-03
Exchange/demand reaction	EX_lac_D(e)_f	D-lactate exchange	-0.26	7.67E-03
Exchange/demand reaction	EX_HC02203(e)_f	prostaglandin-a2 exchange	-0.27	5.65E-03
Exchange/demand reaction	EX_glyc(e)_b	Glycerol exchange	-0.26	7.97E-03
Exchange/demand reaction	EX_his_L(e)_b	exchange reaction for L-histidine	-0.27	5.03E-03
Exchange/demand reaction	EX_met_L(e)_b	L-Methionine exchange	-0.27	5.15E-03
Exchange/demand reaction	EX_pglyc_hs(e)_b	phosphatidylglycerol (homo sapiens) exchange	-0.27	5.03E-03
Exchange/demand reaction	EX_phe_L(e)_b	exchange reaction for L-phenylalanine	-0.27	5.36E-03
Exchange/demand reaction	EX_thr_L(e)_b	L-Threonine exchange	-0.27	5.03E-03
Exchange/demand reaction	EX_trp_L(e)_b	L-Tryptophan exchange	-0.27	5.03E-03
Exchange/demand reaction	EX_xylt(e)_b	exchange reaction for xylitol	-0.30	2.15E-03
Folate metabolism	MTHFC_f	methenyltetrahydrofolate cyclohydrolase	0.27	4.91E-03
Folate metabolism	MTHFD2_f	methylenetetrahydrofolate dehydrogenase (NAD)	0.28	3.38E-03
Fructose and mannose metabolism	r0191	UTP:D-fructose-6-phosphate 1-phosphotransferase EC:2.7.1.11	-0.28	3.41E-03
Glycerophospholipid metabolism	CLS_hs	cardiolipin synthase (homo sapiens)	-0.27	5.03E-03
Glycine, serine, alanine and threonine metabolism	r0160	L-Serine:pyruvate aminotransferase Glycine, serine and threonine metabolism EC:2.6.1.5	-0.25	9.06E-03
Glycine, serine, alanine and threonine metabolism	r0553_f	p-Cumic alcohol:NADP+ oxidoreductase Glycine, serine and threonine metabolism EC:1.2.1.8	-0.27	5.99E-03

Pathway	Reaction	Reaction name	Pearson r	p-value
Glycine, serine, alanine and threonine metabolism	r0552_b	p-cumic alcohol:NAD+ oxidoreductase Glycine, serine and threonine metabolism EC:1.2.1.8	-0.27	5.99E-03
Glycolysis/gluconeogenesis	ENO_f	enolase	-0.27	5.23E-03
Glycolysis/gluconeogenesis	FBA_f	fructose-bisphosphate aldolase	-0.26	7.75E-03
Glycolysis/gluconeogenesis	GAPD_f	glyceraldehyde-3-phosphate dehydrogenase	-0.29	2.51E-03
Glycolysis/gluconeogenesis	TPI_f	triose-phosphate isomerase	-0.30	1.85E-03
Glycolysis/gluconeogenesis	r0165	UTP:pyruvate O2-phosphotransferase EC:2.7.1.40	-0.27	6.42E-03
Glycolysis/gluconeogenesis	PGK_b	phosphoglycerate kinase	-0.29	2.43E-03
Glycolysis/gluconeogenesis	PGM_b	phosphoglycerate mutase	-0.26	6.77E-03
Nucleotide interconversion	CYTK6_f	cytidylate kinase (CMP,CTP)	-0.30	2.26E-03
Pentose phosphate pathway	XYLUR_b	xylulose reductase	-0.27	6.22E-03
Pentose phosphate pathway	r0784_b	xylitol:NAD oxidoreductase Pentose and glucuronate interconversions EC:1.1.1.15	-0.27	5.92E-03
Pyrimidine synthesis	RE0453C_b	RE0453	-0.26	8.24E-03
Pyruvate metabolism	LDH_D_b	D-lactate dehydrogenase	-0.26	7.67E-03
Sphingolipid metabolism	DHCRD1	dihydroceramide desaturase	-0.27	5.28E-03
Sphingolipid metabolism	DSAT	dihydrosphingosine N-acyltransferase	-0.27	5.03E-03
Sphingolipid metabolism	SMS	Sphingomyelin synthase (homo sapiens)	-0.27	5.03E-03
Squalene and cholesterol synthesis	HMGCOARc	Hydroxymethylglutaryl CoA reductase (ir) in cytosol	-0.27	5.03E-03
Transport, endoplasmic reticular	CHSTEROLtrc_f	transport of cholesterol into the cytosol	-0.27	5.03E-03
Transport, endoplasmic reticular	FRDPtcr_f	transport of Farnesyl diphosphate into the endoplasmic reticulum	-0.27	6.55E-03
Transport, endoplasmic reticular	FORtr_b	FOR transporter, endoplasmic reticulum	-0.27	5.03E-03
Transport, endoplasmic reticular	r1051_b	Vesicular transport	-0.27	5.03E-03
Transport, extracellular	ALAGLYexR_f	L-alanine/glycine reversible exchange	0.28	3.90E-03
Transport, extracellular	PGLYCt_f	phosphatidylglycerol transport	-0.27	5.03E-03
Transport, extracellular	XYLTt_f	Xylitol transport via passive diffusion	-0.30	2.15E-03
Transport, extracellular	r1993	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.15	-0.29	3.14E-03
Transport, extracellular	r1996	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.15	-0.25	9.16E-03
Transport, extracellular	r1997	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.15	-0.28	4.13E-03
Transport, extracellular	r2006	Amino Acid-Polyamine-Organocation (APC) TCDB:2.A.3.8.15	-0.26	6.62E-03
Transport, extracellular	r2203_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	-0.25	9.52E-03
Transport, extracellular	r2204_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	-0.25	9.52E-03
Transport, extracellular	r2205_f	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.14	-0.25	9.52E-03
Transport, extracellular	D_LACt2_b	D-lactate transport via proton symport	-0.26	7.67E-03
Transport, extracellular	r2094_b	Major Facilitator(MFS) TCDB:2.A.1.13.1	-0.29	3.20E-03
Transport, extracellular	r2260_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	-0.27	5.19E-03

Pathway	Reaction	Reaction name	Pearson r	p-value
Transport, extracellular	r2261_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	-0.27	5.19E-03
Transport, extracellular	r2262_b	Resistance-Nodulation-Cell Division (RND) TCDB:2.A.60.1.5	-0.27	5.19E-03
Transport, lysosomal	r1052	Vesicular transport	-0.27	5.03E-03
Transport, mitochondrial	r1146_f	Biosynthesis of steroids Enzyme catalyzed	-0.27	5.03E-03
Transport, mitochondrial	CHSTEROLt2_f	cholesterol intracellular transport	-0.27	5.03E-03
Transport, mitochondrial	CHSTEROLt3_b	cholesterol intracellular transport	-0.27	5.03E-03
Transport, nuclear	DGTPtn_f	dGTP diffusion in nucleus	-0.27	5.03E-03
Transport, peroxisomal	FRDPtr_f	lipid, flip-flop intracellular transport	-0.28	4.56E-03

Appendix C

Development of a diagnostic panel for lysosomal storage disorders

LSDs are a group of monogenic metabolic disorders, each one leading to the accumulation of specific substrates due to the deficit of a lysosomal hydrolase. Although individually rare, overall incidence of
acrshortpllsd is estimated around 1:5000-1:8000 [198]. Affected children generally appear normal at birth and the first signs and symptoms develop during the first few months of life and progressively worsen. However LSDs can occur also as late-onset juvenile and adult forms. The diagnosis of LSDs requires clinical expertise as most features are not specific and could be shared by different LSDs; in some cases the diagnosis could be very difficult and may take several years. The first diagnostic assessments are biochemical assays to evaluate the accumulation of specific substrates and/or the enzymatic activity of one or more enzymes. Then molecular analysis of the suspected gene is performed to reveal the disease-causing genetic variants. This diagnostic route could be potentially reversed given the accessibility to NGS technologies which allow the simultaneous sequencing of several genes in a short time. An approach of targeting sequencing could be the primary screening tool in the diagnosis of LSDs, thereby drastically shortening the time from the onset of first symptoms to the diagnosis formulation.

In this study we evaluated a targeted sequencing panel as a potential diagnostic tool for LSDs. In particular, bioinformatics analysis was involved in the panel design and validation. An important novelty is the inclusion of intronic regions, which may allow identifying novel putative regulatory regions. This feature requires hypothesising functionally relevant sequences in non-coding regions and predicting the potentially pathogenic effect of mutations.

C.1 Materials and methods

C.1.1 Selection of target genomic regions

In the selection of target genes, we employed the Orphanet list of LSDs, the Society for the Study of Inborn Errors of Metabolism LSD list and the list reported by Fernandez-Marmiesse and colleagues in their panel design [199]. Genes associated with extremely rare disorders and those disorders presenting a very peculiar phenotype were removed from the list.

We developed a two-step pipeline to obtain the genomic coordinates of target regions, including both exonic and intronic sequences (Figure C.1). First, we selected likely functionally relevant regions by means of an evolutionary approach, in order to identify the CIF [200]. Base-wise conservation estimates were calculated from a multiple alignment through a Hidden Markov model. We chose to align against the genomes of 33 placental mammalian organisms to get reasonably relevant sequences. Obtained regions are merged together and filtered according to their length and distance. Next, we combined Ensembl and RefSeq annotations to define exonic and intronic regions, thereby obtaining all exons and CIFs within the genes of interest [120, 201].

Full list of genes included in the diagnostic panel

AGA	GUSB
ARSA	HEXA
ARSB	HEXB
ASAH1	HGSNAT
CLN3	HYAL1
CLN5	IDS
CLN6	IDUA
CLN8	LAMP2
CTNS	LIPA
CTSA	MAN2B1
CTSD	MANBA
CTSK	MCOLN1
DNAJC5	MFSD8
FUCA1	NAGA
GAA	NAGLU
GALC	NEU1
GALNS	NPC1
GBA	NPC2
GLA	PPT1
GLB1	PSAP
GM2A	SGSH
GNE	SLC17A5
GNPTAB	SMPD1
GNPTG	SUMF1
GNS	TPP1

Table C.1 A total of 50 genes were included in the panel, associated to most LSDs.

The values of filtering parameters were chosen in such a way to optimise coverage and number of sequences relatively to final costs: conservation score greater than 0.85, 20 base pairs minimum length, 2 base pairs maximum distance between two fragments. For each gene the first 50 CIFs with highest score were included in the panel design. The Ion AmpliSeq™ platform (Thermo Fisher Scientific) was then used to design a custom panel including the protein-coding transcripts of selected genes. For each transcript, its exons, a 50 base pairs flanking sequence on each side and both untranslated regions were given to the Ion Ampliseq™ Designer software as target sequence.

Pipeline for the determination of target regions and construction of amplicons

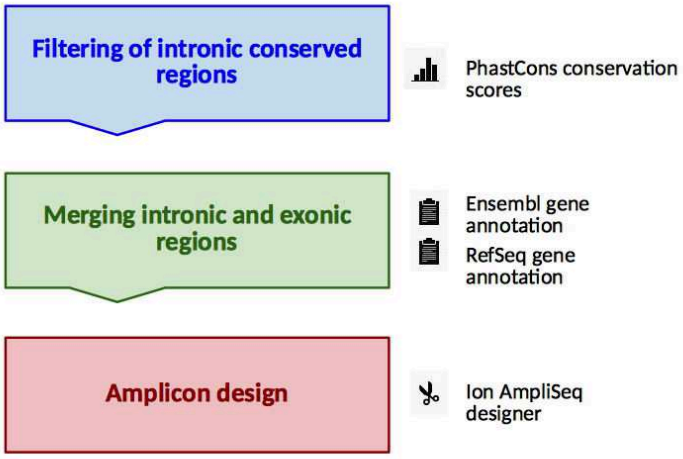


Figure C.1 First, PhastCons conservation scores were used to establish the CIFs. Second, exonic and intronic target regions were determined by means of multiple gene annotations. Third, the amplicons were designed including all coding and non-coding target regions.

C.1.2 Samples selection

In order to validate the panel, a total of 80 samples were collected from different European Clinical and Diagnostic Centers and from the Telethon cell line and DNA Biobank from patients affected by genetic diseases [202]. 59 of them were positive controls, 12 belonged to patients who were diagnosed only through enzymatic analysis and the remaining 9 came from suspected LSD patients for which a diagnosis had not been formulated yet.

C.1.3 Variants analysis

We combined three different variant callers to identify reliable mutations. A first analysis of variants was performed using QueryOR, a platform for variants prioritisation [203]. Prioritisation criteria included frequency less than 0.01, being associated to a true major allele in the reference genome, possessing deleterious types of substitution and various *in silico* pathogenicity scores.

Intronic variants located in the CIFs were filtered in the same manner and further analysed using different tools. Variants falling in regulatory regions and predicted to have a deleterious impact were obtained through Ensembl Variant Effect Predictor (VEP) [204]. SPANR was used to predict both intronic and exonic SNPs affecting RNA splicing [205]. For each variant

up to 300 nucleotides inside an intron, the tool returns a score representing how strongly the variant is predicted to impair exon skipping.

C.2 Results

The total target sequence length was 202.59 kilo bases and included 50 LSD genes (Table C.1) and 230 CIF with an average length of 40 base pairs. The panel design output was a 187.42 kilo bases sequence covered by 1561 amplicons, with an average amplicon length of 240 base pairs and 93% of the whole target sequence covered. Considering only exons, their flanking sequences and untranslated regions, the target sequence coverage was 92.4%. The less covered genes resulted DNAJC5, CLN8, IDUA, NPC2, HYAL 1, whose sequence was covered for a percentage between 55% and 80%. Considering only the coding sequence the most affected gene is IDUA with 8 exons being partially or totally uncovered.

Variant analysis lead to the identification of pathogenic variants in 64% of the positive controls. Failed variant detections are caused by: (i) variants not covered by the amplicons due to panel design; (ii) lowly-covered variants due to poor amplification of specific amplicons; (iii) large deletions not detected by QueryOR. The analysis of CIF focused on those samples from undiagnosed patients whose no variants had been found through the previous analysis. 345 intronic SNPs with frequency less than 0.01 or with no frequency (not annotated variants) filtered by QueryOR were analysed by SPANR and VEP. We selected 14 variants via SPANR as potentially deleterious, but unfortunately none of them were carried by samples from undiagnosed patients. The same intronic variants analysed by VEP gave 61 variants mapping in regulatory regions whose six were carried by undiagnosed samples in promoters, in promoter flanking regions or in enhancers. A deeper analysis of these variants is ongoing to verify through by other tools their potential pathogenicity.

The panel analysis lead to confirmation of previous enzymatic diagnoses for 6 out of 12 subjects in which we found both mutations. In two and three samples respectively we found only one mutation or no mutations. Moreover, two new diagnoses were achieved among the 9 undiagnosed patients. Sanger validations till now performed on biochemically diagnosed samples confirmed the panel results with exception of one case in which a poor covered missense mutation was not revealed in the sample.

Full list of SNPs potentially altering splicing

Variant ID	Chromosome	Position	Reference nucleotide	Alternative nucleotide	Transcript	dPSI	dPSI percentile	Frequency
rs140673721	22	42463768	C	T	chr22:NAGA:-:NM_000262:Exon3	-8,42	0,3	< 0.01
rs113077439	10	73581632	C	T	chr10:PSAP:-:NM_002778:Exon8	-14,93	0,1	0
rs113077439	10	73581632	C	T	chr10:PSAP:-:NM_001042466:Exon8	-17,15	0,09	0
rs193302855	16	1412610	A	G	chr16:GNPTG:+:NM_032520:Exon9	-31,89	0,04	0
rs773281453	17	78079694	G	C	chr17:GAA:+:NM_001079803:Exon4	-22,25	0,06	0
rs773281453	17	78079694	G	C	chr17:GAA:+:NM_000152:Exon3	-22,25	0,06	0
rs773281453	17	78079694	G	C	chr17:GAA:+:NM_001079804:Exon3	-22,25	0,06	0
rs80338815	22	51065593	C	T	chr22:ARSA:-:NM_000487:Exon2	-22,45	0,06	0
rs80338815	22	51065593	C	T	chr22:ARSA:-:NM_001085428:Exon2	-17,57	0,08	0
rs121965020	4	981646	C	T	chr4:IDUA:+:NM_000203:Exon2	-24,96	0,05	0
rs398123224	X	100653456	G	A	chrX:GLA:-:NM_000169:Exon6	-11,75	0,2	0
NA	12	102179788	A	G	chr12:GNPTAB:-:NM_024312:Exon5	-19,16	0,08	NA
NA	1	25254063	A	C	chr1:RUNX3:-:NM_004350:Exon2	-13,39	0,2	NA
NA	14	73678476	G	A	chr14:PSEN1:+:NM_000021:Exon10	-24,55	0,05	NA
NA	14	73678476	G	A	chr14:PSEN1:+:NM_007318:Exon10	-24,55	0,05	NA
NA	17	27975363	C	T	chr17:SSH2:-:NM_001282129:Exon14	-27,93	0,05	NA
NA	17	27975363	C	T	chr17:SSH2:-:NM_033389:Exon13	-27,93	0,05	NA
NA	1	89585861	C	G	chr1:GBP2:-:NM_004120:Exon4	-11,5	0,2	NA
NA	19	7591323	A	G	chr19:MCOLN1:+:NM_020533:Exon3	-14,31	0,2	NA
NA	22	31654275	A	C	chr22:LIMK2:+:NM_016733:Exon2	-10,93	0,2	NA
NA	22	31654275	A	C	chr22:LIMK2:+:NM_001031801:Exon2	-10,93	0,2	NA

Table C.2 Variants predicted by SPANR as potentially damaging exon skipping. NA entries represent non-annotated variants.

Bibliography

[1] T Strachan and A Read. *Human molecular genetics*. Garland Science, New York, NY, USA, 2011.

[2] D Botstein and N Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, 33:22–237, 2003.

[3] 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(suppl 1):D514–D517, 2005. doi:10.1093/nar/gki033.

[4] Charles R. Scriver and Paula J. Waters. Monogenic traits are not simple: lessons from phenylketonuria. *Trends in Genetics*, 15(7):267 – 272, 1999. ISSN 0168-9525. doi:http://dx.doi.org/10.1016/S0168-9525(99)01761-8.

[5] Ganesh Sriram, Julian A. Martinez, Edward R.B. McCabe, James C. Liao, and Katrina M. Dipple. Single-gene disorders: What role could moonlighting enzymes play? *The American Journal of Human Genetics*, 76(6):911 – 924, 2005. ISSN 0002-9297. doi:http://dx.doi.org/10.1086/430799.

[6] Katrina M. Dipple and Edward R.B. McCabe. Modifier genes convert simple mendelian disorders to complex traits. *Molecular Genetics and Metabolism*, 71 (12):43 – 50, 2000. ISSN 1096-7192. doi:https://doi.org/10.1006/mgme.2000.3052.

[7] Katrina M. Dipple and Edward R.B. McCabe. Phenotypes of patients with simple mendelian disorders are complex traits: Thresholds, modifiers, and systems dynamics. *The American Journal of Human Genetics*, 66(6):1729 – 1735, 2000. ISSN 0002-9297. doi:http://dx.doi.org/10.1086/302938.

[8] Jose L. Badano and Nicholas Katsanis. Beyond mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet*, 3(10):779–789, 10 2002.

[9] Michael L. Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, 01 2010.

[10] Fatih Ozsolak and Patrice M. Milos. Rna sequencing: advances, challenges and opportunities. *Nat Rev Genet*, 12(2):87–98, 02 2011.

[11] Peter W. Laird. Principles and challenges of genome-wide dna methylation analysis. *Nat Rev Genet*, 11(3):191–203, 02 2010.

[12] Peter J. Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, 10 2009.

[13] A. F. Maarten Altelaar, Javier Munoz, and Albert J. R. Heck. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet*, 14(1): 35–48, 01 2013.

[14] Vladimir Shulaev. Metabolomics technology and bioinformatics. *Briefings in Bioinformatics*, 7(2):128–139, 2006. doi:10.1093/bib/bbl012.

[15] Solveig K. Sieberts and Eric E. Schadt. Moving toward a system genetics view of disease. *Mammalian Genome*, 18(6):389–401, Jul 2007. ISSN 1432-1777. doi:10.1007/s00335-007-9040-6.

[16] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, May 2017. ISSN 1474-760X. doi:10.1186/s13059-017-1215-1.

[17] Trudy F. C. Mackay, Eric A. Stone, and Julien F. Ayroles. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*, 10(8):565–577, 08 2009.

[18] Mete Civelek and Aldons J. Lusis. Systems genetics approaches to understand complex traits. *Nat Rev Genet*, 15(1):34–48, 01 2014.

[19] Atul J. Butte. Translational bioinformatics: Coming of age. *Journal of the American Medical Informatics Association*, 15(6):709–714, 2008. doi:10.1197/jamia.M2824.

[20] Leroy Hood. Predictive, preventive, participatory, and personalized health. *Rambam Maimonides Med. J.*, 4:e0012, April 2013.

[21] Marylyn D. Ritchie, Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*, 16(2):85–97, 02 2015.

[22] Laurent Heirendt and et al. Creation and analysis of biochemical constraint-based models: the cobra toolbox v3.0. *submitted*, 2017.

[23] Edward J. O'Brien, Jonathan M. Monk, and Bernhard O. Palsson. Using genome-scale models to predict biological capabilities. *Cell*, 161(5):971 – 987, 2015. ISSN 0092-8674. doi:https://doi.org/10.1016/j.cell.2015.05.019.

[24] Jens Nielsen. Systems biology of metabolism: A driver for developing personalized and precision medicine. *Cell Metabolism*, 25(3):572 – 579, 2017. ISSN 1550-4131. doi:https://doi.org/10.1016/j.cmet.2017.02.002.

[25] Daniel R. Hyduke, Nathan E. Lewis, and Bernhard O. Palsson. Analysis of omics data with genome-scale models of metabolism. *Mol. BioSyst.*, 9:167–174, 2013. doi:10.1039/C2MB25453K.

[26] R.P. Vivek-Ananth and Areejit Samal. Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. *Biosystems*, 147:1 – 10, 2016. ISSN 0303-2647. doi:https://doi.org/10.1016/j.biosystems.2016.06.001.

[27] Claudio Angione and Pietro Lió. Predictive analytics of environmental adaptability in multi-omic network models. *Scientific Reports*, 5, 10 2015.

[28] Claudio Angione. Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism. *Bioinformatics*, btx562, 2017. doi:10.1093/bioinformatics/btx562.

[29] Neil Swainston, Kieran Smallbone, Hooman Hefzi, Paul D. Dobson, Judy Brewer, Michael Hanscho, Daniel C. Zielinski, Kok Siong Ang, Natalie J. Gardiner, Jahir M. Gutierrez, Sarantos Kyriakopoulos, Meiyappan Lakshmanan, Shangzhong Li, Joanne K. Liu, Veronica S. Martínez, Camila A. Orellana, Lake-Ee Quek, Alex Thomas, Juergen Zanghellini, Nicole Borth, Dong-Yup Lee, Lars K. Nielsen, Douglas B. Kell, Nathan E. Lewis, and Pedro Mendes. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12(7):109, 2016. ISSN 1573-3890. doi:10.1007/s11306-016-1051-4.

[30] Adil Mardinoglu, Rasmus Agren, Caroline Kampf, Anna Asplund, Mathias Uhlen, and Jens Nielsen. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat Commun*, 5:3083 EP –, 01 2014.

[31] Xiang Zhang, Jan A. Kuivenhoven, and Albert K. Groen. Forward individualized medicine from personal genomes to interactomes. *Frontiers in Physiology*, 6:364, 2015. ISSN 1664-042X. doi:10.3389/fphys.2015.00364.

[32] D Salgado, MI Bellgard, JP Desvignes, and C Béroud. How to identify pathogenic mutations among all those variations: Variant annotation and filtration in the genome sequencing era. *Human Mutation*, 37(12):1272–1282, 2016.

[33] D Altshuler, M Daly, and L Kruglyak. Guilt by association. *Nature Genetics*, 26: 135–137, 2000. doi:doi:10.1038/79839.

[34] Gregory A. Petsko. Guilt by association. *Genome Biology*, 10(4):104, 2009. ISSN 1474-760X. doi:10.1186/gb-2009-10-4-104.

[35] Lin Hou, Min Chen, Clarence K. Zhang, Judy Cho, and Hongyu Zhao. Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Human Molecular Genetics*, 23(10):2780–2790, 2014. doi:10.1093/hmg/ddt668.

[36] Arunachalam Vinayagam, Travis E. Gibson, Ho-Joon Lee, Bahar Yilmazel, Charles Roesel, Yanhui Hu, Young Kwon, Amitabh Sharma, Yang-Yu Liu, Norbert Perrimon, and Albert-László Barabási. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences*, 113(18):4976–4981, 2016. doi:10.1073/pnas.1603992113.

[37] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46:310 EP –, 02 2014.

[38] Max Schubach, Matteo Re, Peter N. Robinson, and Giorgio Valentini. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Scientific Reports*, 7(1):2959, 2017. doi:10.1038/s41598-017-03011-5.

[39] Sahar Gelfman, Quanli Wang, K. Melodi McSweeney, Zhong Ren, Francesca La Carpia, Matt Halvorsen, Kelly Schoch, Fanni Ratzon, Erin L. Heinzen, Michael J. Boland, Slavé Petrovski, and David B. Goldstein. Annotating pathogenic non-coding variants in genic regions. *Nature Communications*, 8(1):236, 2017. doi:10.1038/s41467-017-00141-2.

[40] Arcady R. Mushegian, Douglas E. Bassett, Mark S. Boguski, Peer Bork, and Eugene V. Koonin. Positionally cloned human disease genes: Patterns of evolutionary conservation and functional motifs. *Proceedings of the National Academy of Sciences*, 94(11): 5831–5836, 1997.

[41] 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. doi:10.1093/nar/gkh605.

[42] Zhidong Tu, Li Wang, Min Xu, Xianghong Zhou, Ting Chen, and Fengzhu Sun. Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, 7(1):31, Feb 2006. ISSN 1471-2164. doi:10.1186/1471-2164-7-31.

[43] Tomislav Domazet-Lošo and Diethard Tautz. An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular Biology and Evolution*, 25(12): 2699–2707, 2008. doi:10.1093/molbev/msn214.

[44] Nick G.C. Smith and Adam Eyre-Walker. Human disease genes: patterns and predictions. *Gene*, 318:169 – 175, 2003. ISSN 0378-1119. doi:https://doi.org/10.1016/S0378-1119(03)00772-8.

[45] Fyodor A. Kondrashov, Aleksey Y. Ogurtsov, and Alexey S. Kondrashov. Bioinformatical assay of human gene morbidity. *Nucleic Acids Research*, 32(5):1731–1737, 2004. doi:10.1093/nar/gkh330.

[46] Naoki Osada, Shuhei Mano, and Jun Gojobori. Quantifying dominance and deleterious effect on human disease genes. *Proceedings of the National Academy of Sciences*, 106 (3):841–846, 2009. doi:10.1073/pnas.0810433106.

[47] Katrina M. Dipple, James K. Phelan, and Edward R.B. McCabe. Consequences of complexity within biological networks: Robustness and health, or vulnerability and disease. *Molecular Genetics and Metabolism*, 74(1–2):45 – 50, 2001. ISSN 1096-7192. doi:https://doi.org/10.1006/mgme.2001.3227.

[48] M Oti, B Snel, M A Huynen, and H G Brunner. Predicting disease genes using protein–protein interactions. *Journal of Medical Genetics*, 43(8):691–698, 2006. ISSN 0022-2593. doi:10.1136/jmg.2006.041376.

[49] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68, 01 2011.

[50] Marc Vidal, Michael E. Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986 – 998, 2011. ISSN 0092-8674. doi:http://dx.doi.org/10.1016/j.cell.2011.02.016.

[51] Michael P. H. Stumpf, Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008. doi:10.1073/pnas.0708078105.

[52] Min-Sik Kim, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S. Manda, Raghothama Chaerkady, Anil K. Madugundu, Dhanashree S. Kelkar, Ruth Isserlin, Shobhit Jain, Joji K. Thomas, Babylakshmi Muthusamy, Pamela Leal-Rojas, Praveen Kumar, Nandini A. Sahasrabuddhe, Lavanya Balakrishnan, Jayshree Advani, Bijesh George, Santosh Renuse, Lakshmi Dhevi N. Selvan, Arun H. Patil, Vishalakshi Nanjappa, Aneesha Radhakrishnan, Samarjeet Prasad, Tejaswini Subbannayya, Rajesh Raju, Manish Kumar, Sreelakshmi K. Sreenivasamurthy, Arivusudar Marimuthu, Gajanan J. Sathe, Sandip Chavan, Keshava K. Datta, Yashwanth Subbannayya, Apeksha Sahu, Soujanya D. Yelamanchi, Savita Jayaram, Pavithra Rajagopalan, Jyoti Sharma, Krishna R. Murthy, Nazia Syed, Renu Goel, Aafaque A. Khan, Sartaj Ahmad, Gourav Dey, Keshav Mudgal, Aditi Chatterjee, Tai-Chung Huang, Jun Zhong, Xinyan Wu, Patrick G. Shaw, Donald Freed, Muhammad S. Zahari, Kanchan K. Mukherjee, Subramanian Shankar, Anita Mahadevan, Henry Lam, Christopher J. Mitchell, Susarla Krishna Shankar, Parthasarathy Satishchandra, John T. Schroeder, Ravi Sirdeshmukh, Anirban Maitra, Steven D. Leach, Charles G. Drake, Marc K. Halushka, T. S. Keshava Prasad, Ralph H. Hruban, Candace L. Kerr, Gary D. Bader, Christine A. Iacobuzio-Donahue, Harsha Gowda, and Akhilesh Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, 05 2014.

[53] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber, and Bernhard Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, 05 2014.

[54] Thomas Rolland, Murat Tasan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, Atanas Kamburov, Susan D. Ghiassian, Xinping Yang, Lila Ghamsari, Dawit Balcha, Bridget E. Begg, Pascal Braun, Marc Brehme, Martin P. Broly, Anne-Ruxandra Carvunis, Dan Convery-Zupan, Roser Corominas, Jasmin Coulombe-Huntington, Elizabeth Dann, Matija Dreze, Amélie Dricot, Changyu Fan, Eric Franzosa, Fana Gebreab, Bryan J. Gutierrez, Madeleine F. Hardy, Mike Jin, Shuli Kang, Ruth Kiros, Guan Ning Lin, Katja Luck, Andrew MacWilliams, Jörg Menche, Ryan R. Murray, Alexandre Palagi, Matthew M. Poulin, Xavier Rambout, John Rasla, Patrick Reichert, Viviana Romero, Elien Ruyssinck, Julie M. Sahalie, Annemarie Scholz, Akash A. Shah, Amitabh Sharma, Yun Shen, Kerstin Spirohn, Stanley Tam, Alexander O. Tejeda, Shelly A. Trigg, Jean-Claude Twizere, Kerwin Vega, Jennifer Walsh, Michael E. Cusick, Yu Xia, Albert-László Barabási, Lilia M. Iakoucheva, Patrick Aloy, Javier De Las Rivas, Jan

Tavernier, Michael A. Calderwood, David E. Hill, Tong Hao, Frederick P. Roth, and Marc Vidal. A proteome-scale map of the human interactome network. *Cell*, 159(5): 1212 – 1226, 2014. ISSN 0092-8674. doi:http://dx.doi.org/10.1016/j.cell.2014.10.050.

[55] Edward L. Huttlin, Lily Ting, Raphael J. Bruckner, Fana Gebreab, Melanie P. Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, Rui Dong, Virginia Guarani, Laura Pontano Vaites, Alban Ordureau, Ramin Rad, Brian K. Erickson, Martin Wühr, Joel Chick, Bo Zhai, Deepak Kolippakkam, Julian Mintseris, Robert A. Obar, Tim Harris, Spyros Artavanis-Tsakonas, Mathew E. Sowa, Pietro De Camilli, Joao A. Paulo, J. Wade Harper, and Steven P. Gygi. The bioplex network: A systematic exploration of the human interactome. *Cell*, 162(2):425 – 440, 2015. ISSN 0092-8674. doi:http://dx.doi.org/10.1016/j.cell.2015.06.043.

[56] Edward L. Huttlin, Raphael J. Bruckner, Joao A. Paulo, Joe R. Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P. Gygi, Hannah Parzen, John Szpyt, Stanley Tam, Gabriela Zarraga, Laura Pontano-Vaites, Sharan Swarup, Anne E. White, Devin K. Schweppe, Ramin Rad, Brian K. Erickson, Robert A. Obar, K. G. Guruharsha, Kejie Li, Spyros Artavanis-Tsakonas, Steven P. Gygi, and J. Wade Harper. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, 05 2017.

[57] Mathieu Dalvai, Jeremy Loehr, Karine Jacquet, Caroline C. Huard, Céline Roques, Pauline Herst, Jacques Côté, and Yannick Doyon. A scalable genome-editing-based approach for mapping multiprotein complexes in human cells. *Cell Reports*, 13(3):621 – 633, 2015. ISSN 2211-1247. doi:http://dx.doi.org/10.1016/j.celrep.2015.09.009.

[58] Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng, Li Qiang, Yu Shi, Chan Aye Thu, Brygida Bisikirska, Celine Lefebvre, Domenico Accili, Tony Hunter, Tom Maniatis, Andrea Califano, and Barry Honig. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–560, 10 2012.

[59] Jong Ghut Ashley Aw, Yang Shen, Andreas Wilm, Miao Sun, Xin Ni Lim, Kum Loong Boon, Sidika Tapsin, Yun Shen Chan, Cheng Peow Tan, Adelene Y.L. Sim, Tong Zhang, Teodorus Theo Susanto, Zhiyan Fu, Niranjan Nagarajan, and Yue Wan. In vivo mapping of eukaryotic RNA interactomes reveals principles of higher–order organization and regulation. *Molecular Cell*, 62(4):603 – 617, 2016. ISSN 1097-2765. doi:https://doi.org/10.1016/j.molcel.2016.04.028.

[60] Zhipeng Lu, Qiangfeng Cliff Zhang, Byron Lee, Ryan A. Flynn, Martin A. Smith, James T. Robinson, Chen Davidovich, Anne R. Gooding, Karen J. Goodrich, John S. Mattick, Jill P. Mesirov, Thomas R. Cech, and Howard Y. Chang. RNA duplex map in living cells reveals higher–order transcriptome structure. *Cell*, 165(5):1267 – 1279, 2016. ISSN 0092-8674. doi:https://doi.org/10.1016/j.cell.2016.04.028.

[61] Eesha Sharma, Tim Sterne-Weiler, Dave O'Hanlon, and Benjamin J. Blencowe. Global mapping of human RNA–RNA interactions. *Molecular Cell*, 62(4):618 – 626, 2016. ISSN 1097-2765. doi:https://doi.org/10.1016/j.molcel.2016.04.030.

[62] Maria D. Paraskevopoulou, Ioannis S. Vlachos, Dimitra Karagkouni, Georgios Georgakilas, Ilias Kanellos, Thanasis Vergoulis, Konstantinos Zagganas, Panayiotis Tsanakas, Evangelos Floros, Theodore Dalamagas, and Artemis G. Hatzigeorgiou.

Diana-lncbase v2: indexing microrna targets on non-coding transcripts. *Nucleic Acids Research*, 44(D1):D231–D238, 2016. doi:10.1093/nar/gkv1270.

[63] TM Cafarelli, A Desbuleux, Y Wang, SG Choi, D De Ridder, and M Vidal. Mapping, modeling, and characterization of protein-protein interactions on a proteomic scale. *Current Opinion in Structural Biology*, 44:201 – 210, 2017. ISSN 0959-440X. doi:http://dx.doi.org/10.1016/j.sbi.2017.05.003.

[64] Alessandro Ori, Murat Iskar, Katarzyna Buczak, Panagiotis Kastritis, Luca Parca, Amparo Andrés-Pons, Stephan Singer, Peer Bork, and Martin Beck. Spatiotemporal variation of mammalian protein complex stoichiometries. *Genome Biology*, 17(1):47, Mar 2016. ISSN 1474-760X. doi:10.1186/s13059-016-0912-5.

[65] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D. Spear, Carolyn S. Sevier, Huiming Ding, Judice L.Y. Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P. St. Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J. Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L. Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li, Zhen-Yuan Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph San Luis, Ermira Shuteriqi, Amy Hin Yan Tong, Nydia van Dyk, Iain M. Wallace, Joseph A. Whitney, Matthew T. Weirauch, Guoqing Zhong, Hongwei Zhu, Walid A. Houry, Michael Brudno, Sasan Ragibizadeh, Balázs Papp, Csaba Pál, Frederick P. Roth, Guri Giaever, Corey Nislow, Olga G. Troyanskaya, Howard Bussey, Gary D. Bader, Anne-Claude Gingras, Quaid D. Morris, Philip M. Kim, Chris A. Kaiser, Chad L. Myers, Brenda J. Andrews, and Charles Boone. The genetic landscape of a cell. *Science*, 327(5964):425–431, 2010. ISSN 0036-8075. doi:10.1126/science.1180823.

[66] Cecily J. Wolfe, Isaac S. Kohane, and Atul J. Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6(1):227, Sep 2005. ISSN 1471-2105. doi:10.1186/1471-2105-6-227.

[67] Sipko vanDam, Thomas Craig, and João Pedro deMagalh aes. Genefriends: a human rna-seq-based gene and transcript co-expression database. *Nucleic Acids Research*, 43 (D1):D1124, 2015. doi:10.1093/nar/gku1042.

[68] Noam Auslander, Allon Wagner, Matthew Oberhardt, and Eytan Ruppin. Data-driven metabolic pathway compositions enhance cancer survival prediction. *PLOS Computational Biology*, 12(9):1–17, 09 2016. doi:10.1371/journal.pcbi.1005125.

[69] Laurence D Hurst. It's easier to get along with the quiet neighbours. *Molecular Systems Biology*, 13(9), 2017. doi:10.15252/msb.20177961.

[70] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015. doi:10.1093/nar/gku1179.

[71] Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5): 610 – 615, 2008. ISSN 0002-9297. doi:http://dx.doi.org/10.1016/j.ajhg.2008.09.017.

[72] 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, Dec 2004. ISSN 1474-760X. doi:10.1186/gb-2004-6-1-r7.

[73] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012. doi:10.1093/nar/gkr988.

[74] Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, Lisa Matthews, Bruce May, Marija Milacic, Karen Rothfels, Veronica Shamovsky, Marissa Webber, Joel Weiser, Mark Williams, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487, 2016. doi:10.1093/nar/gkv1351.

[75] Atanas Kamburov, Ulrich Stelzl, Hans Lehrach, and Ralf Herwig. The consensuspathdb interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D793–D800, 2013. doi:10.1093/nar/gks1055.

[76] 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(suppl 1):D5–D16, 2010. doi:10.1093/nar/gkp967.

[77] Kwang-Il Goh and In-Geol Choi. Exploring the human diseasome: the human disease network. *Briefings in Functional Genomics*, 11(6):533–542, 2012. doi:10.1093/bfgp/els032.

[78] Marc A van Driel, Jorn Bruggeman, Gert Vriend, Han G Brunner, and Jack A M Leunissen. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5): 535–542, 2006. doi:10.1038/sj.ejhg.5201585.

[79] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004. doi:10.1093/nar/gkh061.

[80] LJ Jensen, M Kuhn, M Stark, S Chaffron, C Creevey, J Muller, T Doerks, P Julien, A Roth, M Simonovic, P Bork, and C von Mering. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, 37(Database issue):D412–D416, 2009.

[81] Marc A. van Driel and Han G. Brunner. Bioinformatics methods for identifying candidate disease genes. *Human Genomics*, 2(6):429, Jun 2006. ISSN 1479-7364. doi:10.1186/1479-7364-2-6-429.

[82] Maricel G. Kann. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Briefings in Bioinformatics*, 11(1):96–110, 2010. doi:10.1093/bib/bbp048.

[83] Rosario M. Piro and Ferdinando Di Cunto. Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS Journal*, 279(5):678–696, 2012. ISSN 1742-4658. doi:10.1111/j.1742-4658.2012.08471.x.

[84] Y Moreau and LC Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet*, 13:523–536, 2012.

[85] W. Verleyen, S. Ballouz, and J. Gillis. Measuring the wisdom of the crowds in network-based gene function inference. *Bioinformatics*, 31(5):745–752, 2015. doi:10.1093/bioinformatics/btu715.

[86] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nat Rev Genet*, 16(6):321–332, 06 2015.

[87] M. K. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey. Machine learning in genomic medicine: A review of computational problems and data sets. *Proceedings of the IEEE*, 104(1):176–197, Jan 2016. ISSN 0018-9219. doi:10.1109/JPROC.2015.2494198.

[88] S Aerts, D Lambrechts, S Maity, P Van Loo, B Coessens, F De Smet, LC Tranchevent, Bart De Moor, P Marynen, B Hassan, P Carmeliet, and Y Moreau. Gene prioritization through genomic data fusion. *Nat Biotech*, 24(5):537–544, 2006.

[89] J Shawe-Taylor and N Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, New York, NY, USA, 2004.

[90] M Gönen and E Alpaydın. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, 2011.

[91] X Wang, EP Xing, and DJ Schaid. Kernel methods for large-scale genomic data analysis. *Briefings in Bioinformatics*, 16(2):183–192, 2015.

[92] KM. Borgwardt, CS Ong, S Schönauer, SVN Vishwanathan, AJ Smola, and HP Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl 1):i47–i56, 2005.

[93] T De Bie, LC Tranchevent, LMM van Oeffelen, and Y Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–i132, 2007.

[94] S Yu, T Falck, A Daemen, LC Tranchevent, JAK Suykens, B De Moor, and Y Moreau. L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, 11(1):309, 2010.

[95] F Mordelet and JP Vert. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, 12(1): 389, 2011.

[96] P Zakeri, S Elshal, and Y Moreau. Gene prioritization through geometric-inspired kernel data fusion. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1559–1565, 2015.

[97] O Chapelle, B Schölkopf, and A Zien. *Semi-supervised learning*. MIT Press, Cambridge, MA, USA, 2006.

[98] Xiaotu Ma, Ting Chen, and Fengzhu Sun. Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. *Briefings in Bioinformatics*, 15(5):685–698, 2014. doi:10.1093/bib/bbt041.

[99] Giorgio Valentini, Alberto Paccanaro, Horacio Caniza, Alfonso E. Romero, and Matteo Re. An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artificial Intelligence in Medicine*, 61 (2):63 – 78, 2014. ISSN 0933-3657. doi:https://doi.org/10.1016/j.artmed.2014.03.003.

[100] B Chen, M Li, J Wang, X Shang, and FX Wu. A fast and high performance multiple data integration algorithm for identifying human disease genes. *BMC Medical Genomics*, 8(3):S2, 2015.

[101] Tianshu Yin, Shu Chen, Xiaohui Wu, and Weidong Tian. GenePANDA–a novel network-based gene prioritizing tool for complex diseases. *Scientific Reports*, 7:43258 EP –, 03 2017.

[102] Léon-Charles Tranchevent, Francisco Bonachela Capdevila, Daniela Nitsch, Bart De Moor, Patrick De Causmaecker, and Yves Moreau. A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1):22–32, 2011. doi:10.1093/bib/bbq007.

[103] F Aiolli and M Donini. Easymkl: a scalable multiple kernel learning algorithm. *Neurocomputing*, 169:215–224, 2015.

[104] François Fouss, Kevin Francoisse, Luh Yen, Alain Pirotte, and Marco Saerens. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural Networks*, 31(Supplement C):53 – 72, 2012. ISSN 0893-6080. doi:https://doi.org/10.1016/j.neunet.2012.03.001.

[105] RI Kondor and JD Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002.

[106] B Chen, M Li, J Wang, and FX Wu. Disease gene identification by using graph kernels and markov random fields. *Science China Life Sciences*, 57(11):1054–1063, 2014.

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

[108] P Chebotarev and E Shamis. The matrix-forest theorem and measuring relations in small social groups. *Automation and Remote Control*, 58(9):1505–1514, 1997.

[109] F Aiolli, G Da San Martino, and A Sperduti. A kernel method for the optimization of the margin distribution. In *Proceedings of the ICANN Conference, Praga 2008*, pages 16–27, 2008.

[110] M Frasca, JF Fontaine, G Valentini, M Mesiti, M Notaro, D Malchiodi, and MA Andrade-Navarro. Disease genes must guide data source integration in the gene prioritization process. In *The 14th International Conference on Bioinformatics and Biostatistics (CIBB)*, 2017.

[111] M Polato and F Aiolli. Kernel based collaborative filtering for very large scale top-n item recommendation. In *Proceedings of the ESANN Conference, Bruges 2016*, 2016.

[112] PA Devijver and J Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, London, GB, 1982.

[113] KI Goh, ME Cusick, D Valle, B Childs, M Vidal, and AL Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[114] JA Hanley and BJ McNeil. The meaning and the use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

[115] 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(suppl 1): D767–D772, 2009. doi:10.1093/nar/gkn892.

[116] C Wu, C Orozco, J Boyer, M Leglise, J Goodale, S Batalov, CL Hodge, J Haase, J Janes, JW Huss, and AI Su. Biogps: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome biology*, 10(11):R130, 2009.

[117] M Whirl-Carrillo, EM McDonagh, JM Hebert, L Gong, K Sangkuhl, CF Thorn, RB Altman, and TE Klein. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther*, 92(4):414–417, 2012.

[118] CF Schaefer, K Anthony, S Krupa, J Buchoff, M Day, T Hannay, and KH Buetow. Pid: the pathway interaction database. *Nucleic Acids Research*, 37(Database Issue): D674–D679, 2008.

[119] D Börnigen, LC Tranchevent, F Bonachela-Capdevila, K Devriendt, B De Moor, P De Causmaecker, and Y Moreau. An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23):3081–3088, 2012.

[120] Bronwen L. Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón,

Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin, Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek, and Stephen M. J. Searle. The ensembl gene annotation system. *Database*, 2016:baw093, 2016. doi:10.1093/database/baw093.

[121] Jesse Gillis and Paul Pavlidis. Guilt by association is the exception rather than the rule in gene networks. *PLOS Computational Biology*, 8(3):1–13, 03 2012. doi:10.1371/journal.pcbi.1002444.

[122] Ralf Bender and Stefan Lange. Adjusting for multiple testing–when and how? *Journal of Clinical Epidemiology*, 54(4):343 – 349, 2001. ISSN 0895-4356. doi:https://doi.org/10.1016/S0895-4356(00)00314-0.

[123] Yixuan Chen, Wenhui Wang, Yingyao Zhou, Robert Shields, Sumit K. Chanda, Robert C. Elston, and Jing Li. In silico gene prioritization by integrating multiple data sources. *PloS One*, 6(6):e21137, 2011.

[124] S Köhler, S Bauer, D Horn, and PN Robinson. Walking the interactome for prioritization of candidate disease genes. *American journal of human genetics*, 82(4):949–958, 2008.

[125] D Seelow, JM Schwarz, and M Schuelke. Genedistiller-distilling candidate genes from linkage intervals. *PLoS One*, 3(12):e3874, 2008.

[126] J Chen, H Xu, BJ Aronow, and AG Jegga. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, 8(1):392, 2007.

[127] D Nitsch, JP. Gonçalves, F Ojeda, B de Moor, and Y Moreau. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11(1):460, 2010.

[128] JE Hutz, AT Kraja, HL McLeod, and MA Province. Candid: a flexible method for prioritizing candidate genes for complex human traits. *Genet Epidemiol*, 32(8):816, 2008.

[129] EA Adie, RR Adams, KL Evans, DJ Porteous, and BS Pickard. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22(6):773–774, 2006.

[130] Y Yoshida, Y Makita, N Heida, S Asano, A Matsushima, M Ishii, Y Mochizuki, H Masuya, S Wakana, N Kobayashi, and T 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–W152, 2009.

[131] Stefanie Weber. Novel genetic aspects of congenital anomalies of kidney and urinary tract. *Curr Opin Pediatr*, 24(8):212–8, 2012. doi:10.1097/MOP.0b013e32834fdbd4.

[132] Pawaree Saisawat, Velibor Tasic, Virginia Vega-Warner, Elijah O. Kehinde, Barbara Günther, Rannar Airik, Jeffrey W. Innis, Bethan E. Hoskins, Julia Hoefele, Edgar A. Otto, and Friedhelm Hildebrandt. Identification of two novel cakut-causing genes by massively parallel exon resequencing of candidate genes in patients with unilateral renal agenesis. *Kidney International*, 81(2):196 – 200, 2012. ISSN 0085-2538. doi:https://doi.org/10.1038/ki.2011.315.

[133] Hila Barak, Sung-Ho Huh, Shuang Chen, Cécile Jeanpierre, Jelena Martinovic, Mélanie Parisot, Christine Bole-Feysot, Patrick Nitschké, Rémi Salomon, Corinne Antignac, David M. Ornitz, and Raphael Kopan. Fgf9 and fgf20 maintain the stemness of nephron progenitors in mice and man. *Developmental Cell*, 22(6):1191–1207, 2017. doi:10.1016/j.devcel.2012.04.018.

[134] Shan Elahi, Alison Homstad, Himani Vaidya, Jennifer Stout, Gentzon Hall, Guanghong Wu, Peter Conlon, Jonathan C. Routh, John S. Wiener, Sherry S. Ross, Shashi Nagaraj, Delbert Wigfall, John Foreman, Adebowale Adeyemo, Indra R. Gupta, Patrick D. Brophy, C. Egla Rabinovich, and Rasheed A. Gbadegesin. Rare variants in tenascin genes in a cohort of children with primary vesicoureteric reflux. *Pediatric Nephrology*, 31(2):247–253, Feb 2016. ISSN 1432-198X. doi:10.1007/s00467-015-3203-6.

[135] Stefanie Weber, Jaclyn C. Taylor, Paul Winyard, Kari F. Baker, Jessica Sullivan-Brown, Raphael Schild, Tanja Knüppel, Aleksandra M. Zurowska, Alberto Caldas-Alfonso, Mieczyslaw Litwin, Sevinc Emre, Gian Marco Ghiggeri, Aysin Bakkaloglu, Otto Mehls, Corinne Antignac, Escape Network, Franz Schaefer, and Rebecca D. Burdine. Six2 and bmp4 mutations associate with anomalous kidney development. *Journal of the American Society of Nephrology*, 19(5):891–903, 2008. doi:10.1681/ASN.2006111282.

[136] Kansuporn Sriyudthsak, Fumihide Shiraishi, and Masami Yokota Hirai. Mathematical modeling and dynamic simulation of metabolic reaction systems using metabolome time series data. *Frontiers in molecular biosciences*, 3:15, 2016. ISSN 2296-889X. doi:10.3389/fmolb.2016.00015.

[137] Ina Aretz and David Meierhofer. Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *International Journal of Molecular Sciences*, 17(5), 2016. ISSN 1422-0067. doi:10.3390/ijms17050632.

[138] JD Orth, I Thiele, and BO Palsson. What is flux balance analysis? *Nat Biotech*, 28: 245–248, 2010.

[139] NE Lewis, H Nagarajan, and BO Palsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Micro*, 10: 291–305, 2012.

[140] Ali Ebrahim, Elizabeth Brunk, Justin Tan, Edward J O'Brien, Donghyuk Kim, Richard Szubin, Joshua A Lerman, Anna Lechner, Anand Sastry, Aarash Bordbar, Adam M Feist, and Bernhard O Palsson. Multi-omic data integration enables discovery of hidden biological regularities. *Nat Commun*, 7:13091, 2016. doi:10.1038/ncomms13091.

[141] Willi Gottstein, Brett G. Olivier, Frank J. Bruggeman, and Bas Teusink. Constraint-based stoichiometric modelling from single organisms to microbial communities. *Journal of The Royal Society Interface*, 13(124), 2016. ISSN 1742-5689. doi:10.1098/rsif.2016.0627.

[142] Jeremy S. Edwards and Bernhard O. Palsson. Systems properties of the Haemophilus influenzaeRd metabolic genotype. *Journal of Biological Chemistry*, 274(25):17410–17416, 1999. doi:10.1074/jbc.274.25.17410.

[143] J. S. Edwards and B. O. Palsson. The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10):5528–5533, 2000. doi:10.1073/pnas.97.10.5528.

[144] Natalie C. Duarte, Scott A. Becker, Neema Jamshidi, Ines Thiele, Monica L. Mo, Thuy D. Vo, Rohith Srivas, and Bernhard O. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007. doi:10.1073/pnas.0610772104.

[145] Daniel J Cook and Jens Nielsen. Genome-scale metabolic models applied to human health and disease. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, pages e1393–n/a, 2017. ISSN 1939-005X. doi:10.1002/wsbm.1393. e1393.

[146] Hamideh Fouladiha and Sayed-Amir Marashi. Biomedical applications of cell- and tissue-specific metabolic network models. *Journal of Biomedical Informatics*, 68:35 – 49, 2017. ISSN 1532-0464. doi:http://dx.doi.org/10.1016/j.jbi.2017.02.014.

[147] Nathan Lewis and Alyaa Abdel-Haleem. The evolution of genome-scale models of cancer metabolism. *Frontiers in Physiology*, 4:237, 2013. ISSN 1664-042X. doi:10.3389/fphys.2013.00237.

[148] Itay Shaked, Matthew A. Oberhardt, Nir Atias, Roded Sharan, and Eytan Ruppin. Metabolic network prediction of drug side effects. *Cell Systems*, 2(3):209–213, 2018/01/14 . doi:10.1016/j.cels.2016.03.001.

[149] Minseung Kim, Navneet Rai, Violeta Zorraquino, and Ilias Tagkopoulos. Multi-omics integration accurately predicts cellular state in unexplored conditions for escherichia coli. *Nature Communications*, 7:13090 EP –, 10 2016.

[150] Tomer Shlomi, Tomer Benyamini, Eyal Gottlieb, Roded Sharan, and Eytan Ruppin. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect. *PLOS Computational Biology*, 7(3):1–8, 03 2011. doi:10.1371/journal.pcbi.1002018.

[151] Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, 13(4):227–232, 04 2012. doi:doi:10.1038/nrg3185.

[152] Jingyi Jessica Li, Peter J. Bickel, and Mark D. Biggin. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, 2:e270, February 2014. ISSN 2167-8359. doi:10.7717/peerj.270.

[153] Marko Jovanovic, Michael S. Rooney, Philipp Mertins, Dariusz Przybylski, Nicolas Chevrier, Rahul Satija, Edwin H. Rodriguez, Alexander P. Fields, Schraga Schwartz, Raktima Raychowdhury, Maxwell R. Mumbach, Thomas Eisenhaure, Michal Rabani, Dave Gennert, Diana Lu, Toni Delorey, Jonathan S. Weissman, Steven A. Carr, Nir Hacohen, and Aviv Regev. Dynamic profiling of the protein life cycle in response to pathogens. *Science*, 347(6226), 2015. ISSN 0036-8075. doi:10.1126/science.1259038.

[154] Idit Kosti, Nishant Jain, Dvir Aran, Atul J. Butte, and Marina Sirota. Cross-tissue analysis of gene and protein expression in normal and cancer tissues. *Scientific Reports*, 6:24799 EP –, 05 2016.

[155] Ehsan Motamedian, Maryam Mohammadi, Seyed Abbas Shojaosadati, and Mona Heydari. Trfba: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. *Bioinformatics*, 33(7): 1057, 2017. doi:10.1093/bioinformatics/btw772.

[156] Joao C. Guimaraes, Miguel Rocha, and Adam P. Arkin. Transcript level and sequence determinants of protein abundance and noise in escherichia coli. *Nucleic Acids Research*, 42(8):4791–4799, 2014. doi:10.1093/nar/gku126.

[157] Marius Paltanea, Sabin Tabirca, Ernesc Scheiber, and Mark Tangney. Logarithmic growth in biological processes. In *Proceedings of the 2010 12th International Conference on Computer Modelling and Simulation*, UKSIM '10, pages 116–121, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4016-0. doi:10.1109/UKSIM.2010.29.

[158] Helena Firczuk, Shichina Kannambath, Jürgen Pahle, Amy Claydon, Robert Beynon, John Duncan, Hans Westerhoff, Pedro Mendes, and John EG McCarthy. An in vivo control map for the eukaryotic mrna translation machinery. *Molecular Systems Biology*, 9(1), 2013. doi:10.1038/msb.2012.73.

[159] Kazuyuki Shimizu. *Metabolic Flux Analysis Based on 13C-Labeling Experiments and Integration of the Information with Gene and Protein Expression Patterns*, pages 1–49. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. doi:10.1007/b94204.

[160] Steffen Klamt, Georg Regensburger, Matthias P. Gerstl, Christian Jungreuthmayer, Stefan Schuster, Radhakrishnan Mahadevan, Jürgen Zanghellini, and Stefan Müller. From elementary flux modes to elementary flux vectors: Metabolic pathway analysis with arbitrary linear flux constraints. *PLOS Computational Biology*, 13(4):1–22, 04 2017. doi:10.1371/journal.pcbi.1005409.

[161] Daniel C. Zielinski, Neema Jamshidi, Austin J. Corbett, Aarash Bordbar, Alex Thomas, and Bernhard O. Palsson. Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism. *Scientific Reports*, 7:41241 EP –, 01 2017.

[162] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, Karl K Weitz, Roland Eils, Rainer König, Richard D Smith, and Bernhard Ø Palsson. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6(1), 2010. ISSN 1744-4292. doi:10.1038/msb.2010.47.

[163] Sharon J. Wiback, Iman Famili, Harvey J. Greenberg, and Bernhard Palsson. Monte carlo sampling can be used to determine the size and shape of the steady-state flux space. *Journal of Theoretical Biology*, 228(4):437 – 447, 2004. ISSN 0022-5193. doi:https://doi.org/10.1016/j.jtbi.2004.02.006.

[164] Nathan D. Price, Jan Schellenberger, and Bernhard O. Palsson. Uniform sampling of steady-state flux spaces: Means to design experiments and to interpret enzymopathies. *Biophysical Journal*, 87(4):2172 – 2186, 2004. ISSN 0006-3495. doi:https://doi.org/10.1529/biophysj.104.043000.

[165] Alfredo Braunstein, Anna Paola Muntoni, and Andrea Pagnani. An analytic approximation of the feasible space of metabolic networks. *Nat Commun*, 8:14915 EP –, 04 2017.

[166] JK. et al White. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell*, 154(2):452–464, 2013. doi:10.1016/j.cell.2013.06.022.

[167] Nielsen J Robinson JL. Integrative analysis of human omics data using biomolecular networks. *Mol Biosyst*, 12:2953–64, 2016.

[168] Thomas N. Seyfried, Roberto E. Flores, Angela M. Poff, and Dominic P. D'Agostino. Cancer as a metabolic disease: implications for novel therapeutics. *Carcinogenesis*, 35(3):515–527, 2014. doi:10.1093/carcin/bgt480.

[169] Robert H. Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer*, 6(10):813–823, 10 2006.

[170] Christian Diener and Osbaldo Resendis-Antonio. Personalized prediction of proliferation rates and metabolic liabilities in cancer biopsies. *Frontiers in Physiology*, 7:644, 2016. ISSN 1664-042X. doi:10.3389/fphys.2016.00644.

[171] Omer F. Kuzu, Mohammad A. Noory, and Gavin P. Robertson. The role of cholesterol in cancer. *Cancer Research*, 76(8):2063–2070, 2016. ISSN 0008-5472. doi:10.1158/0008-5472.CAN-15-2613.

[172] Natalya N. Pavlova and Craig B. Thompson. The emerging hallmarks of cancer metabolism. *Cell Metabolism*, 23(1):27–47, 2017. doi:10.1016/j.cmet.2015.12.006.

[173] Arkaitz Carracedo, Lewis C. Cantley, and Pier Paolo Pandolfi. Cancer metabolism: fatty acid oxidation in the limelight. *Nat Rev Cancer*, 13(4):227–232, 04 2013.

[174] S. Rodríguez-Enríquez, L. Hernández-Esquivel, A. Marín-Hernández, M. El Hafidi, J.C. Gallardo-Pérez, I. Hernández-Reséndiz, J.S. Rodríguez-Zavala, S.C. Pacheco-Velázquez, and R. Moreno-Sánchez. Mitochondrial free fatty acid oxidation supports oxidative phosphorylation and proliferation in cancer cells. *International Journal of Biochemistry and Cell Biology*, 65:209–221, 2015. doi:10.1016/j.biocel.2015.06.010.

[175] Fokhrul Hossain, Amir A. Al-Khami, Dorota Wyczechowska, Claudia Hernandez, Liqin Zheng, Krzystoff Reiss, Luis Del Valle, Jimena Trillo-Tinoco, Tomasz Maj, Weiping Zou, Paulo C. Rodriguez, and Augusto C. Ochoa. Inhibition of fatty acid

oxidation modulates immunosuppressive functions of myeloid-derived suppressor cells and enhances cancer therapies. *Cancer Immunology Research*, 3(11):1236–1247, 2015. ISSN 2326-6066. doi:10.1158/2326-6066.CIR-15-0036.

[176] Kathryn E. Hopperton, Robin E. Duncan, Richard P. Bazinet, and Michael C. Archer. Fatty acid synthase plays a role in cancer metabolism beyond providing fatty acids for phospholipid synthesis or sustaining elevations in glycolytic activity. *Experimental Cell Research*, 320(2):302 – 310, 2014. ISSN 0014-4827. doi:http://dx.doi.org/10.1016/j.yexcr.2013.10.016.

[177] Kevin J Rycyna, Dean J Bacich, and Denise S O'Keefe. Opposing roles of folate in prostate cancer. *Urology*, 82(6):1197–1203, December 2013. ISSN 0090-4295. doi:10.1016/j.urology.2013.07.012.

[178] Andrzej Stepulak, Radoslaw Rola, Krzysztof Polberg, and Chrysanthy Ikonomidou. Glutamate and its receptors in cancer. *Journal of Neural Transmission*, 121(8):933–944, Aug 2014. ISSN 1435-1463. doi:10.1007/s00702-014-1182-6.

[179] V. Dolce, A.R. Cappello, R. Lappano, and M. Maggiolini. Glycerophospholipid synthesis as a novel drug target against cancer. *Current Molecular Pharmacology*, 4 (3):167–175, 2011. doi:10.2174/1874467211104030167.

[180] P M Tedeschi, E K Markert, M Gounder, H Lin, D Dvorzhinski, S C Dolfi, L L-Y Chan, J Qiu, R S DiPaola, K M Hirshfield, L G Boros, J R Bertino, Z N Oltvai, and A Vazquez. Contribution of serine, folate and glycine metabolism to the atp, nadph and purine requirements of cancer cells. *Cell Death Dis*, 4:e877–, 10 2013.

[181] Alberto Chiarugi, Christian Dolle, Roberta Felici, and Mathias Ziegler. The nad metabolome – a key determinant of cancer cell biology. *Nat Rev Cancer*, 12(11): 741–752, 11 2012.

[182] C. Cantó, K.J. Menzies, and J. Auwerx. Nad+ metabolism and the control of energy homeostasis: A balancing act between mitochondria and the nucleus. *Cell Metabolism*, 22(1):31–53, 2015. doi:10.1016/j.cmet.2015.05.023.

[183] Yusuf A. Hannun and Lina M. Obeid. The ceramide-centric universe of lipid-mediated cell regulation: Stress encounters of the lipid kind. *Journal of Biological Chemistry*, 277(29):25847–25850, 2002. doi:10.1074/jbc.R200008200.

[184] Yan Jiang, Yong Pan, Patrea R. Rhea, Lin Tan, Mihai Gagea, Lorenzo Cohen, Susan M. Fischer, and Peiying Yang. A sucrose-enriched diet promotes tumorigenesis in mammary gland in part through the 12-lipoxygenase pathway. *Cancer Research*, 76 (1):24–29, 2016. doi:10.1158/0008-5472.CAN-14-3432.

[185] Bing Wang, Gerd Bobe, John J. LaPres, and Leslie D. Bourquin. High sucrose diets promote intestinal epithelial cell proliferation and tumorigenesis in apc min mice by increasing insulin and igf-i levels. *Nutrition and Cancer*, 61(1):81–93, 2008. doi:10.1080/01635580802372609.

[186] Sejal Vyas, Elma Zaganjor, and Marcia C. Haigis. Mitochondria and cancer. *Cell*, 166 (3):555–566, 2017. doi:10.1016/j.cell.2016.07.002.

[187] Deuk Kim Nam, Eunok Im, Hyun Yoo Young, and Hyun Choi Yung. Modulation of the cell cycle and induction of apoptosis in human cancer cells by synthetic bile acids. *Current Cancer Drug Targets*, 6(8):681–689, 2006. ISSN 1568-0096/1873-5576. doi:10.2174/156800906779010236.

[188] PR Baker, JC Wilton, CE Jones, DJ Stenzel, N Watson, and GJ Smith. Bile acids influence the growth, oestrogen receptor and oestrogen-regulated proteins of mcf-7 human breast cancer cells. *Br J Cancer*, 65(4):566–572, 04 1992.

[189] S. Pavlides, I. Vera, R. Gandara, S. Sneddon, R.G. Pestell, I. Mercier, U.E. Martinez-Outschoorn, D. Whitaker-Menezes, A. Howell, F. Sotgia, and M.P. Lisanti. Warburg meets autophagy: Cancer-associated fibroblasts accelerate tumor growth and metastasis via oxidative stress, mitophagy, and aerobic glycolysis. *Antioxidants and Redox Signaling*, 16(11):1264–1284, 2012. doi:10.1089/ars.2011.4243.

[190] Paola Giussani, Cristina Tringali, Laura Riboni, Paola Viani, and Bruno Venerando. Sphingolipids: Key regulators of apoptosis and pivotal players in cancer drug resistance. *International Journal of Molecular Sciences*, 15(3):4356–4392, 2014. doi:10.3390/ijms15034356.

[191] Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016 01 28 run. 2016. doi:https://doi.org/10.7908/C11G0KM9.

[192] Min Zhao, Pora Kim, Ramkrishna Mitra, Junfei Zhao, and Zhongming Zhao. Tsgene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Research*, 44(D1):D1023–D1031, 2016. doi:10.1093/nar/gkv1268.

[193] Yining Liu, Jingchun Sun, and Min Zhao. Ongene: A literature-based database for human oncogenes. *Journal of Genetics and Genomics*, 44(2):119 – 121, 2017. ISSN 1673-8527. doi:https://doi.org/10.1016/j.jgg.2016.12.004.

[194] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I. Furlong. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833, 2017. doi:10.1093/nar/gkw943.

[195] Y Zhao, H Li, S Fang, Y Kang, W Wu, Y Hao, Z Li, D Bu, N Sun, MQ Zhang, and R Chen. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Research*, 44(D1):D203–D208, 2016.

[196] Gang Luo. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Health Information Science and Systems*, 4 (1):2, Mar 2016. ISSN 2047-2501. doi:10.1186/s13755-016-0015-4.

[197] Guido Zampieri, Dinh Van Tran, Michele Donini, Nicoló Navarin, Fabio Aiolli, Alessandro Sperduti, and Giorgio Valle. Scuba: scalable kernel-based gene prioritization. *BMC Bioinformatics. Accepted*.

[198] Meikle PJ, Hopwood JJ, Clague AE, and Carey WF. Prevalence of lysosomal storage disorders. *JAMA*, 281(3):249–254, 1999. doi:10.1001/jama.281.3.249.

[199] Ana Fernández-Marmiesse, Marcos Morey, Merce Pineda, Jesús Eiris, Maria Luz Couce, Manuel Castro-Gago, Jose Maria Fraga, Lucia Lacerda, Sofia Gouveia, Maria Socorro Pérez-Poyato, Judith Armstrong, Daisy Castiñeiras, and Jose A. Cocho. Assessment of a targeted resequencing assay as a support tool in the diagnosis of lysosomal storage disorders. *Orphanet Journal of Rare Diseases*, 9(1):59, Apr 2014. doi:10.1186/1750-1172-9-59.

[200] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005. doi:10.1101/gr.3715005.

[201] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016. doi:10.1093/nar/gkv1189.

[202] Mirella Filocamo, Raffaella Mazzotti, Fabio Corsolini, Marina Stroppiano, Giorgia Stroppiana, Serena Grossi, Susanna Lualdi, Barbara Tappino, Federica Lanza, Sara Galotto, and Roberta Biancheri. Cell line and dna biobank from patients affected by genetic diseases. *Open Journal of Bioresources*, (1):e2, 2014. doi:10.5334/ojb.ab.

[203] Loris Bertoldi, Claudio Forcato, Nicola Vitulo, Giovanni Birolo, Fabio De Pascale, Erika Feltrin, Riccardo Schiavon, Franca Anglani, Susanna Negrisolo, Alessandra Zanetti, Francesca D'Avanzo, Rosella Tomanin, Georgine Faulkner, Alessandro Vezzi, and Giorgio Valle. Queryor: a comprehensive web platform for genetic variant analysis and prioritization. *BMC Bioinformatics*, 18(1):225, Apr 2017. doi:10.1186/s12859-017-1654-4.

[204] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17(1):122, Jun 2016. doi:10.1186/s13059-016-0974-4.

[205] Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K.C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, and Brendan J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 2014. doi:10.1126/science.1254806.