Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by
M.Sc. Aurelien Jean Fernand Dugourd

born in: 12/04/1989
Oral examination: 31st of January 2022

Bridging the gap between signalling and metabolism through functional and mechanistic analysis of multi-omic data

Referees: Prof. Dr. Julio Saez-Rodriguez
Prof. Dr. Thomas Höfer

Acknowledgement

After highschool, I went through a bachelor and biotechnology master degree without any strong intrinsic motivation. At the end of my biological master thesis, I found myself confronted with a tough choice. I was presented with a thesis opportunity in the direct continuity of my master thesis project, but I knew that I had real interest to continue working in this field.
Yet, As far as I can remember, a big part of my mind was always stuck in the digital world. As an adult, I have probably spent more time in front of a monitor than anywhere else combined. Thus I decided to shift my career path and start over a new master degree, this time in bioinformatic. For this, I will always be grateful to my parents, who actually agreed to financially support me for two more years as a student, while I was turning down opportunities for a funded PhD position.
This was feeling like a huge gamble for me at the time, but it apparently paid off. Through my two years of bioinformatic master degree, my perception of academia and research changed dramatically. As I was discovering that I could actually combine my love for the digital world with an actual socially useful skill, i.e. biological data analysis and modelling, I felt a motivation like I never had before. Studies were not a burden anymore, and in two years I went from being a barely average student to the top student of my master degree's promotion. At this point, I knew that my life would be tied to data analysis forever.
My master degree ended with a lab rotation of 6 month at the Institut Curie in Paris, where I was mentored by Dr. Laurence Calzone. I am very grateful to her, as through her mentoring she allowed me to apply my newly learned skills in real case scenarios. She made me realise for the first time how important it was to factor in the human aspect of the scientific world. She also introduced me to Pr. Julio Saez-Rodriguez, who would then become my PhD thesis supervisor.
Pr. Saez-Rodriguez will always be a scientific role model I look up to. I cannot begin to express my gratitude for his mentoring. He always encouraged my scientific creativity and integrity. He provided me with a work environment that made it possible for my burgeoning skills at the time to bloom to their full potential. Through his example, he taught me that supervisors can be there to actually support and help you, rather than being bosses who just expect obedience and docility. I am also very grateful to all the members of the lab that I worked with over my thesis. I would also like to acknowledge our administrative assistant, Erika Shulz, whose help was extremely precious alongside my thesis.
At the end of my thesis, I also became a part of Pr Kramann's lab. Pr. Rafael Kramann has been of great help to me during my thesis and showed me yet another example of a great and human principal investigator (PI). I also had rotations in four laboratories during the course of my PhD, which allowed me to meet four other amazing PIs : Pr. Christian Frezza, Pr. Miguel Rocha, Pr. Vassily Hatzimanikatis and Pr. Athanasios Mantalaris. I am very grateful for the opportunity they gave me to visit their lab and further broaden my scientific field.
Finally, I am very grateful to Mu-En Chung, whom I met during my thesis and dramatically helped me in learning how to see things through other people's point of view, a skill that can

have much more impact on the quality of our scientific work than I would have initially guessed.

Abstract

In this thesis, I study how biological prior knowledge and high throughput biological data can be systematically integrated to yield mechanistic biological insights. I focused the scope of my work mainly on signalling pathways and metabolism, especially how these two biological functions interact and control each other. The overall goal of this work is to better characterise the molecular driver of complex diseases and chronic health conditions such as cancer, metabolic syndromes and fibrosis. Indeed, if we can better and more systematically understand these conditions, we may be able to design better, more targeted treatments and even prevent them more efficiently.
In the first chapter, I draw a state of the art of multi-omic data generation and how to analyze them in mechanistic contexts. What we call omic data are datasets where the abundance of hundred to thousand unique biological molecules are measured in parallel. Then, in the second chapter, I present a collection of scientific studies where I could learn and apply the principles detailed in the first chapter. In the third chapter, I present my attempt at developing a way to systematically analyse and integrate multiple types of omic data together. The resulting tool, named COSMOS, is presented in the context of a kidney cancer study using multiple types of omic data generated from a cohort of patients. In the final chapter, I present a tool called ocEAn, which aims at estimating metabolic enzyme activity changes from metabolomic data.

Zusammenfassung

In dieser Arbeit untersuche ich, wie biologisches Wissen und biologische Hochdurchsatzdaten systematisch integriert werden können, um mechanistische Erkenntnisse zu gewinnen. In meiner Arbeit konzentrierte ich mich hauptsächlich auf Signalwege und Stoffwechsel, insbesondere wie diese beiden Vorgänge interagieren und sich gegenseitig kontrollieren. Das übergeordnete Ziel dieser Arbeit ist es, den molekularen Treiber komplexer Krankheiten und chronischer Gesundheitszustände wie Krebs, metabolische Syndrome und Fibrose besser zu charakterisieren. Wenn wir diese Erkrankungen besser und systematischer verstehen, können wir möglicherweise bessere und gezieltere Behandlungen entwickeln und ihnen sogar effizienter vorbeugen.
Im ersten Kapitel beschreibe ich den Stand der Technik der Multi-Omic-Datengenerierung und deren Analyse in mechanistischen Zusammenhängen. Was wir Omic-Daten nennen, sind Datensätze, in denen die Häufigkeit von hunderten bis tausenden einzigartigen biologischen Molekülen parallel gemessen wird. Im zweiten Kapitel präsentiere ich dann eine Sammlung wissenschaftlicher Studien, in denen ich die Prinzipien des ersten Kapitels angewandt habe. Im dritten Kapitel beschreibe ich meinen Versuch, einen Weg zur systematischen Analyse und Integration mehrerer Arten von Omic-Daten zu entwickeln. Das resultierende Tool mit dem Namen COSMOS wird im Kontext einer Nierenkrebs Studie vorgestellt, bei der mehrere Arten von Omic-Daten verwendet werden, die von einer Patientenkohorte generiert wurden. Im letzten Kapitel präsentiere ich ein Tool namens ocEAn, das darauf abzielt, Veränderungen der Enzymaktivität aus metabolischen Daten abzuschätzen.

Chapter 1 : From pathways to mechanistic insights; state of the art of multi-omic data analysis and integration

Chapter 1 is a preliminary version of a review that was published in Current opinion in Systems Biology : Footprint-based functional analysis of multi-omic data (Dugourd & Saez-Rodriguez, 2019a). It was written solely by A. Dugourd.

Abstract

Omic technologies allow us to generate extensive data, including transcriptomic, proteomic, phosphoproteomic and metabolomic. These data can be used to study signal transduction, gene regulation and metabolism. In this review, we summarize resources and methods to analysis these types of data. We focus on methods developed to recover functional insights using footprints. Footprints are signatures defined by the effect of molecules or processes of interest. They integrate information from multiple measurements whose abundances are under the influence of a common regulator. For example, transcripts controlled by a transcription factor or peptides phosphorylated by a kinase. Footprints can also be generalised across multiple types of omic data. Thus, we also present methods to integrate multiple types of omic data and features (such as the ones derived from footprints) together. We highlight some examples of studies that leverage such approaches to discover new biological mechanisms.

1. Introduction

In a cell, numerous molecules are constantly interacting and reacting to adapt to the environment and preserve homeostasis. These molecules can be separated in distinct classes, mostly DNA, RNA of various natures (messenger RNA, microRNA, etc,…), proteins and metabolites. They can be subjected to various chemical modifications such as methylation, phosphorylation, ubiquitinilation or glycosylation. Each of these modifications can affect the physical properties of these molecules and, consequently, their functions. In particular, modifications of proteins are often organised in cascades. These cascades are interlinked, forming a complex network that controls most cellular functions. Over the past decades, subparts of this network have been characterized and defined according to the types of reactions and molecules interacting together, notably signaling pathways, regulatory networks, and metabolic networks. Roughly, signaling and regulatory networks represent subnetworks composed mainly of kinases, phosphatases and transcription factors (TFs) connecting proteic sensors (such as membrane receptors) to gene expression. Kinases are responsible for the phosphorylation of proteins while TFs, which are also interconnected, regulate the abundance of RNA transcripts. Metabolic networks are mainly composed of small molecules (metabolites) that are transformed into one another through reactions catalyzed by metabolic enzymes (Figure 1 A and B). Thus, changes in the abundance of phosphorylated proteins, transcripts and metabolites hold information about the functional states of signaling, regulatory and metabolic networks, respectively.

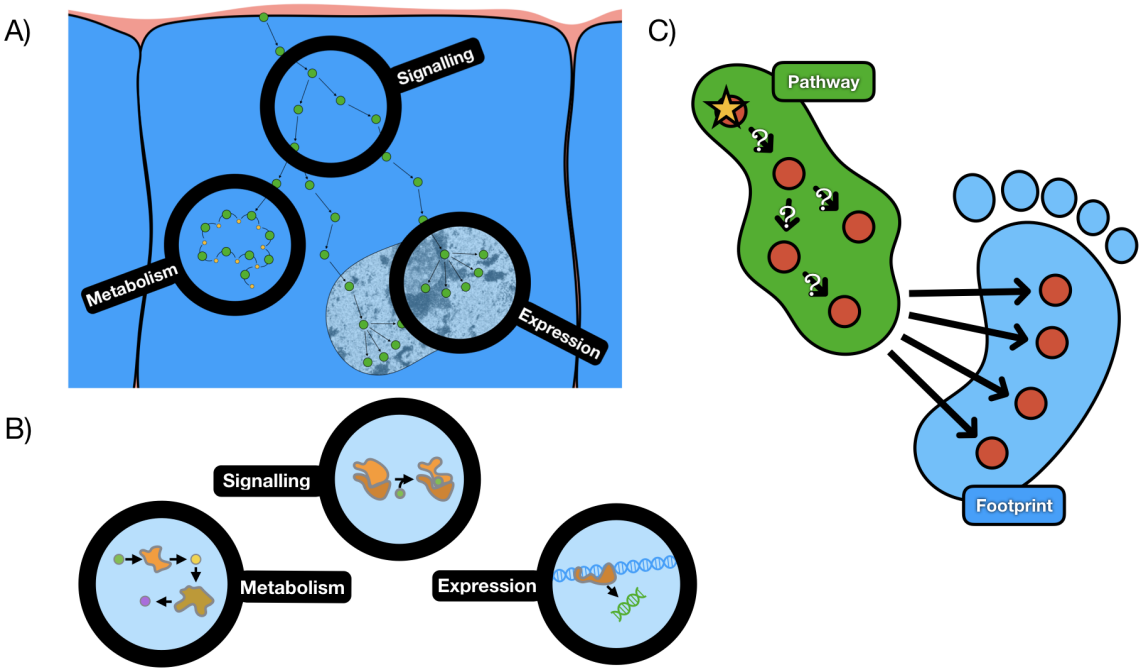


Figure 1 - From pathway to footprint for functional analysis of omic data.
(A,B) Schematic representation of the interactions between signaling, gene regulation and metabolism. The main type of omic data to study them are highlighted. (C) A certain pathway (green) and the potential footprint of perturbing this pathway (blue). The question marks represent the uncertainty of the functionality of interaction in the pathway in a specific context.

Today, it is possible to measure the abundance of thousands of RNA transcripts, protein peptides (chemically modified or not) and metabolites. Such datasets, along with the systematic characterization of other biomolecules (e.g. lipidomics, genomics), are referred to as omic datasets. All these abundances can be considered as the molecular signature of a biological sample in a specific condition, for example cells treated with an enzymatic inhibitor. This concept can also be scaled down at the level of specific enzymes, such as transcription factors or kinases: the abundances of the target transcripts of a transcription factor (TF) can be viewed as the footprint of the TF activity. The same concept applies to the target phospho-peptides of a kinase. A footprint can also be derived for a pathway or process and inform us on their activity. In a classic 'mapping' strategy, the activity of a pathway is inferred from measurements of its own components and the activity of enzymes is estimated from measurements of their corresponding transcripts/proteins. In contrast, footprints based strategies estimate activities from molecular readouts considered to be downstream of the pathway/enzyme (Figure 1 C).

In this review, we will cover recent methods to analyze and extract relevant functional and mechanistic information using molecular signatures applied to omic data. We will also present strategies to integrate together multiple types of omic data. We will focus mainly on molecular measurements directly related to signaling pathways and metabolic reaction networks that can be obtained from transcriptomic, (phospho)proteomic and metabolomic data. We leave out of the scope of the review other omic data, in particular (epi)genomic. Accordingly, we will describe features derived from this data, in particular using footprints. First, we present different types of online knowledge databases that can be used to extract functional insights from omic datasets. Then we summarize mapping and footprint methods, as well as network-based approaches. Finally, we will discuss how these methods can be used to integrate together different types of omic datasets.

2. Prior knowledge resources

2.1 Ontologies and protein-protein interaction databases

A powerful strategy for the analysis of any omic dataset is to integrate it with the current knowledge of the underlying biology. This knowledge is available in multiple resources(Miryala *et al*, 2018), see Table 1. For example, Gene Ontology(Ashburner *et al*, 2000) (GO) is arguably the most used resource for gene annotation. These annotations are very useful to quickly get an overview of molecular functions, cellular compartments and biological processes associated with specific genes. Many other type of annotations, such as signaling pathways, cancer hallmarks, chemical and genetic perturbation signatures, are available in databases such as MSigDB(Liberzon *et al*, 2011). Large resources for protein-protein interactions (PPI) are also available (Miryala *et al*, 2018). For example, STRINGdb(Szklarczyk *et al*, 2015) pulls together many different sources of PPI, from experimentally validated interactions to automatic literature search, while Omnipath(Türei *et al*, 2016) focuses on databases of curated interactions.

Table 1 : Selected Prior knowledge resources discussed in this review.

Database	Content	Link
Brenda	Metabolic enzyme/substrate interactions, reaction networks and enzyme structures.	https://www.brenda-enzymes.org/
CophosK	Kinase/substrate interaction inference.	http://compbio.case.edu/omics/software/cophosk/
Gene Ontology	Molecular functions, biological processes and cellular components	http://geneontology.org/
KEA2	Kinase/substrate interactions from multiple resources.	http://www.maayanlab.net/KEA2/index.html
KEGG	Metabolic enzyme/substrate interactions and reaction networks.	https://www.genome.jp/kegg/
Kinomexplorer	Kinase/substrate interaction inference.	http://kinomexplorer.info/
MSigDB	Gene sets of hallmarks, positions, pathways and perturbation signatures, motifs, gene ontology, oncogenic and immunologic.	http://software.broadinstitute.org/gsea/msigdb

Omnipath	Protein-protein interactions pulled from various resources (Mainly curated). Kinase/substrate interactions. Transcription factor/target interactions (DoRothEA).	http://omnipathdb.org/
Pathway commons	Signaling and metabolic pathways from various databases.	http://www.pathwaycommons.org/
PTMSigDB	Post-translational modification signatures.	https://github.com/broadinstitute/ssGSEA2.0
Reactome	Metabolic enzyme/substrate interactions and reaction networks.	https://reactome.org/
STITCHdb	Chemical/proteins interactions.	http://stitch.embl.de/
STRIBGdb	Protein-protein interactions pulled from various resources (curated and inferred).	https://string-db.org/
Transfac	Transcription factor/target interactions. (Commercial)	http://gene-regulation.com/pub/databases.html
TRRUST	Transcription factor/target interactions.	https://www.grnpedia.org/trrust/

2.2 Enzyme/substrate databases

Databases that capture relationships between enzymes and their substrates are useful to extract relevant information about enzymes from transcriptomic and phosphoproteomic data. These relationships are either predicted with computational methods or experimentally validated.

Transcription factor (TF) targets are available in databases like TRANSFAC(Matys *et al*, 2006) or TRRUST(Han *et al*, 2018). TRRUST uses consensus sequence pattern search to infer potential TF targets, and some of these interactions may be experimentally validated. Hence, the level of confidence in a TF-target interaction can vary. DoRothEA(Garcia-Alonso *et al*, 2018a), which is also embedded in Omnipath(Türei *et al*, 2016), integrates multiple transcription factor target resources (including TRRUST). DoRothEA annotates TF-target interactions with a confidence index based on the source of the interaction (pattern search, experimental validation, etc...). Higher confidence interactions such as experimentally

validated ones seems to yield better estimations of transcription factor activity (Garcia-Alonso *et al*, 2018a).

Similar databases exist for kinases. PhosphositePlus(Hornbeck *et al*, 2012) contains curated information about phosphosites such as their function and kinase/substrate interactions. PTMSigDB(Krug *et al*, 2018), a database of post translational modification signatures combines consensual perturbation footprint across thousands of phosphoproteomic datasets, curated kinase targets and pathways. KinomeExplorer(Horn *et al*, 2014) infers substrate of kinases with amino-acid pattern search and known PPIs. CophosK(Ayati *et al*, 2018) complements experimentally validated databases with correlated phosphosite based on phosphoproteomic data, thus creating context specific kinase/substrate networks. KEA2(Lachmann & Ma'ayan, 2009) and Omnipath(Türei *et al*, 2016) combine together multiple databases of kinase/substrate interactions.

Finally ,information on metabolic enzymes and their targeted metabolites exists in resources such as KEGG(Kanehisa & Goto, 2000), Brenda(Jeske *et al*, 2019), Reactome(Fabregat *et al*, 2018) and REcon3D(Brunk *et al*, 2018).

2.3 Multi-level interaction databases

Some multi-level interaction databases (spanning across multiple different biological processes) already exist. STITCH(Szklarczyk *et al*, 2016), a complement of STRING, combines interactions between chemicals and proteins with PPIs. Omnipath combines TF/targets, kinase/substrate, PPIs and drugs. Pathway Commons combines signaling and metabolic pathways from various databases(Cerami *et al*, 2011). In the future, it is likely that more databases that combine together multiple types of molecular interactions will appear. As more multi-omic datasets are generated, the importance of such combinations of resources will increase.

3. Gene set and pathway enrichment analysis

Gene sets are groups of genes that share a common characteristic (for example, genes that participate in the same biological process). These are available in annotation resources described in 2. Prior knowledge resources. Gene sets can be analysed using multiple methods that can be largely classified as either over-representation or enrichment analysis. Over-representation analysis (ORA) usually tries to answer the following question: when

comparing genes differentially expressed between two conditions, are there sets of genes that contain significantly more differentially expressed genes than expected? Statistical enrichment analysis (EA; often referred to as GSEA-like approaches), tries to answer a slightly different question: when comparing genes differentially expressed between two conditions, are there some sets in which the overall difference of expression is more extreme than expected? EA approaches do so by summarising measurement-level statistics (e.g. fold-changes, t-values, p-values) belonging to the same group/set into a single score and estimate if this summarised score is significantly more extreme than expected (Figure 2 A, see 4. Footprint analysis for a concrete example and (Ackermann & Strimmer, 2009)). While they answer slightly different questions, EA has the advantage that it doesn't require to decide a-priori which genes are significantly changed or not. DAVID(Huang *et al*, 2008) is widely used to run gene set analysis using ORA with GO. Gene Set Enrichment Analysis (GSEA)(Subramanian *et al*, 2005) and Parametric Analysis of Gene Set Enrichment (PAGE)(Kim & Volsky, 2005) are examples of statistical enrichment analysis tools. EnrichR(Kuleshov *et al*, 2016) is a popular platform that provides an intuitive user interface to perform gene set analysis with ORA or EA methods. These tools can also be used with pathway ontologies such as the one present in MSigDB(Liberzon *et al*, 2011) in order to perform pathway enrichment analysis (Figure 2 B). Recent developments in EA take advantage of the underlying topology of pathway. This is done either in a data-driven manner based on correlation between measurements of the same set (Alhamdoosh *et al*, 2017) or using prior knowledge of interactions between members of a pathway (Amadoz *et al*, 2018).

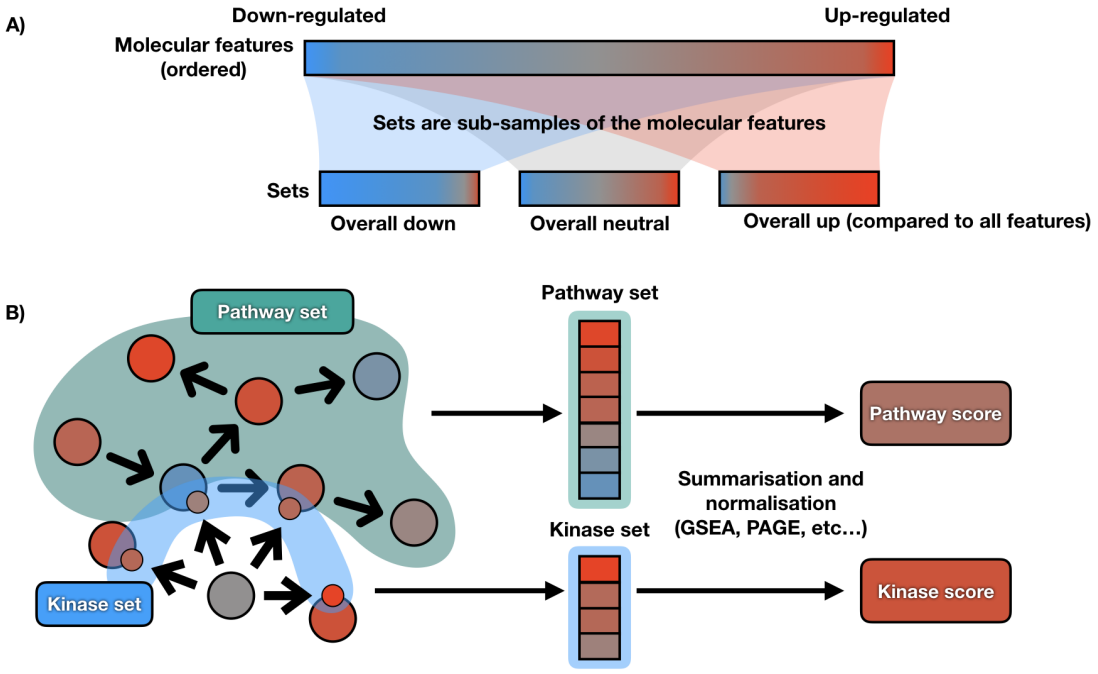


Figure 2 - Comparison between pathway and kinase enrichment analysis

(A) Simplified representation of the fundamental idea of statistical enrichment analysis. Pathways, gene annotation and enzyme targets are sets of molecular features. The goal of an enrichment analysis is to characterise the significance of an overall change of each set compared to the rest of all measured molecular features in a specific condition. (B) In a classic pathway enrichment analysis, the features used to compute the enrichment scores are the members of the pathway itself. In contrast, a kinase enrichment analysis computes the enrichment score with targets of the kinase, but not the kinase itself. The same principle applies for transcription factor and pathway footprint enrichment analysis.

Originally, gene set/pathway enrichment analysis was mainly used to assess whether a specific gene annotation is significantly enriched with extremely deregulated genes. However, this method is very flexible and can be adapted for many different uses. For example, associations between drugs and their expression signature (such as those found in LINCS L1000(Subramanian *et al*, 2017) and DSigDB(Yoo *et al*, 2015)) can be used to identify and repurpose drugs with transcriptome and/or proteomic data.

4. Footprint analysis

EA approaches can also be used for footprint analysis, such as transcription factor and kinase enrichment analysis. Even though the algorithm is the same as for pathway enrichment analysis, the prior knowledge sources are sets of enzyme-targets, fundamentally changing the interpretation and usefulness of enrichment scores. This is possible because, in the case of EA approaches, the enrichment score of a given set directly summarises the changes of the members of the set. Thus, an enrichment score obtained from a set of functional targets of an enzyme can be interpreted directly as a proxy of the activity of this enzyme (Figure 2 B). An example of the procedure to estimate the activity of a kinase with statistical enrichment is shown in Figure 3.

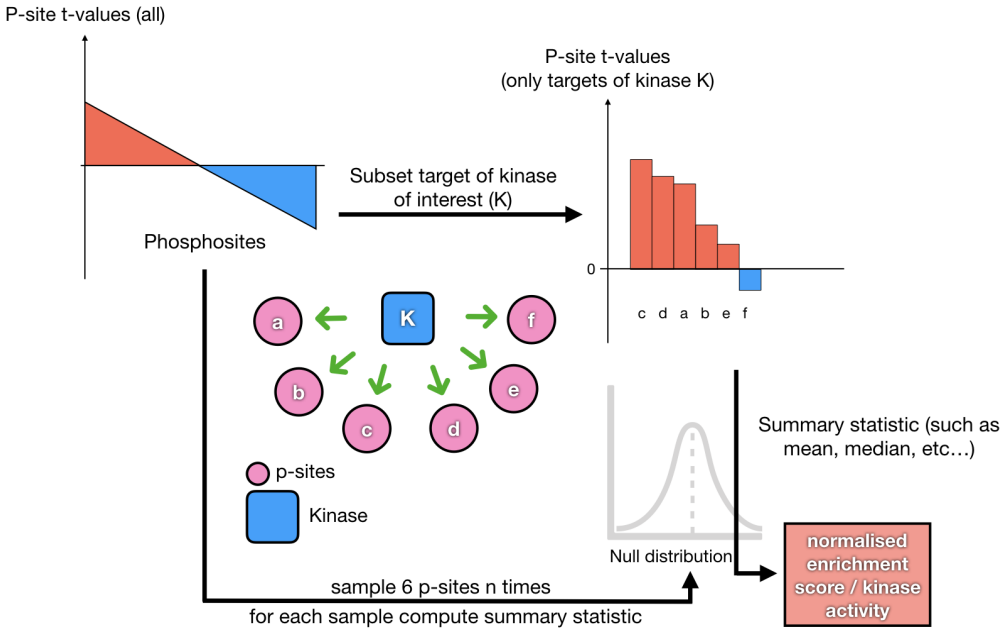


Figure 3 - Example of kinase activity estimation with statistical enrichment analysis

Consider an experiment where the changes in phosphosite abundance were measured between two specific conditions. Given a kinase K that can phosphorylate six phosphosites (a, b, c, d, e, f), one could assume that the changes in abundance of the six phosphosites mirror changes in the activity of kinase K. To estimate this change of activity, the statistics (t-values in this example) associated with the change of abundance of the six targets of kinase K are summarised (using e.g. mean or variance). This summary statistic is called the enrichment score. Then, we need to estimate whether this enrichment score is significantly different from what would be expected from any given set of six phosphosites. To this end, six phosphosites are sampled randomly n times from all the phosphosites available in this study to generate a null distribution of enrichment scores. The enrichment score of kinase K is then normalised with this distribution. Thus, the resulting normalised enrichment score represents how extreme the change in the activity of kinase K is compared to possible kinases randomly associated to phosphosites.

4.1 Transcription factor activity

VIPER is an enrichment analysis method building up on Parametric Analysis of Gene Set Enrichment (PAGE). VIPER can estimate the activity of proteins, typically transcription factors, using the abundance changes of their targets as a proxy of their activity(Alvarez *et al*, 2016). Originally, VIPER uses data-driven inferred TF-targets interactions, but any type of set collection can be used, and it has been applied to the DoRothEA TF-targets interactions mentioned above(Garcia-Alonso *et al*, 2018b).

Osmanbeyoglu et al.(Osmanbeyoglu *et al*, 2017) developed an approach based on bilinear regression to estimate the activity of transcription factors with transcriptomic and phosphoproteomic. This approach directly accounts for phosphorylation events measured upstream of transcription factors when estimating their activity change.

4.2 Kinase activity

Analogously to TFs, the activity of kinases can be estimated from the abundance changes of their substrates from phospho-proteomic data. As for TFs, different statistical models can be used, for example KSEA (Wiredja *et al*, 2017; Casado *et al*, 2013) or KinasePA (Yang *et al*, 2016), an approach specifically tailored to handle datasets with more than two conditions. In (Hernandez-Armenta *et al*, 2017), kinase activity change estimations obtained from various statistical models were compared with kinases knock-out and ligand perturbation datasets. It was shown in this context that simple statistics of the footprint can displayed slightly better agreement with experimental data than more complex statistics such as GSEA or multivariate linear regression models. Yet, the quality of the target set collection seemed to be the main determinant of performance.

4.3 Pathway activity

Tools presented in 3. Gene set and pathway enrichment analysis can yield insight about the activity of pathways using gene expression data (Lim *et al*, 2018). However, these approaches remain limited by the fact that the expression of a gene only partially correlates with the activity of the corresponding protein in a pathway (Krawczenko *et al*, 2017). This limits the amount of information that can be retrieved about the functional state of a pathway from expression measurements related to the members of the pathway itself. An alternative approach is to estimate the activity of the pathway by looking at the genes that are known to change when the pathway is activated or inhibited, akin the footprint methods for kinases and TFs. PROGENy(Schubert *et al*, 2018) (an extension of SPEED(Parikh *et al*, 2010)) learns transcriptomic footprints of a specific pathway from multiple experiments where the pathway is perturbed. Such footprints represent indirect targets downstream of the pathway . They can then be used with the same algorithms presented in 4.1 Transcription factor activity and 4.2 Kinase activity. These footprint genesets have been shown to be more informative than the mapping/ontology genesets (Cantini *et al*, 2018)(Schubert *et al*, 2018).

5. Multi-scale networks

5.1 Correlation-based methods for multi-omic integration

Joint analysis of omic datasets allow us to study the interactions between biological processes. The arguably simplest and most intuitive approach is to use correlation-based methods: correlation between different omic measurements across samples suggests that the processes reflected in one omic regulate the processes reflected by the other, or that there is co-regulation by a third (often unknown) process. This makes it possible to reconstruct networks of interactions based on correlations between multiple measurements and features. For example, correlations between metabolite and metabolic enzyme transcript abundance was estimated in (Auslander *et al*, 2016). This enabled to find mRNA predictors of metabolic abundances. The predicted abundances of these metabolites were, in turn, good predictors of cancer patient survival. A combination of Principal Component Analysis and partial correlation was also used to systematically find pairs of metabolites that are coregulated by either transcriptional or post-transcriptional mechanisms (Schwahn & Nikoloski, 2018). MOFA is a method that generalises Principal Component Analysis to handle multiple omic data(Argelaguet *et al*, 2018). The method was originally applied on a dataset including somatic mutations, RNA expression and DNA methylation, but is in principle applicable to other type of omic datasets such as proteomic, phosphoproteomic and metabolomic and their corresponding footprints (e.g. kinase and transcription factor activities).

Indeed, correlation based approaches can also be used downstream of footprint analysis to connect activity scores with other measurements. For example, kinase activities estimated from phosphoproteomics were correlated with metabolites to find kinases that regulate the activity of metabolic enzymes through post translational modifications (Gonçalves *et al*, 2017).

5.2 Network contextualisation

Most network resources (such as the ones presented in 2. Prior knowledge resources) are generic. They recapitulate all known interactions between omic data in different organisms. However, not all proteins are expressed in all types of cell. Different mutational backgrounds, specially in cancer, can also alter the properties of proteins, such as enzymatic activity and binding ability. Thus, various tools exist to contextualize networks according to specific

conditions (Chen *et al*, 2014)(Tényi *et al*, 2016). These methods combine protein interactions and omic datasets to find significantly deregulated subsets of a larger interaction network. They usually rely on a static protein-protein interaction network and graph theory. Alternative approaches find the most coherent subnetwork connecting perturbation targets (i.e. known proteins that are altered in some way) with deregulated transcripts (Melas *et al*, 2015; Bradley & Barrett, 2017). To do so, protein networks are abstracted as causal models, where nodes (proteins) and edges (interactions) can be active or not. Then, the signed subnetworks that lead to the best fit between its output and experimental measurements are identified. A similar approach was also used in the context of phosphoproteomic data to reconstruct signaling pathways from a generic kinase/substrate network (Terfve *et al*, 2015; Köksal *et al*, 2018). The pathways reconstructed in this way often share similarities with canonical pathways. However, since they use generic prior knowledge networks, they can include nodes that are usually absent from canonical pathways.

In the future, it is likely that such approaches will be generalised to directly integrate multiple type of omic measurements at the same time, combining both measurements and/or output of footprint analysis. In fact, there are already a few examples of recent methods to contextualise networks with multiple type of omic data. The prize-collecting Steiner forest algorithm has been used to find optimal subnetworks in a prior combination of PPI and reaction network based on metabolic and protein abundance measurements (Pirhaji *et al*, 2016). The TieDIE (Drake *et al*, 2016) algorithm can contextualise signaling pathways with specific types of cancer based on transcriptomic and phosphoproteomic data. A pipeline developed by Huna et al. (Huan *et al*, 2018) first extracts relevant metabolic pathways based on metabolomic data and then overlays proteomic and transcriptomic data on these subnetworks. Finally, the HotNet (Reyna *et al*, 2018) algorithm generalises approaches based on graph theory (Chen *et al*, 2014) to find altered subnetwork across multiple biological scales and integrate different types of omic data together.

6. Multi-omic network to find potential actionable treatment targets

To conclude, we believe that integrating multiple types of omic data together using biological knowledge and appropriate computational models will allow us to better understand cellular mechanisms in many contexts (Figure 4). Novel types of regulatory mechanisms in *E. coli* have been discovered by integrating genomic, transcriptomic, ribosomal profiling, proteomic and metabolomic data (Ebrahim *et al*, 2016). Global network reprogramming events occuring

in diabetes have been studied by simultaneously looking at transcriptomic, proteomic, phosphoproteomic and metabolomic changes over a time course (Kawata *et al*, 2018). Post-translation regulatory mechanisms in Fumarate hydratase deficient cancer cells were decoded by integrating proteomic, phosphoproteomic and metabolomic data together (Gonçalves *et al*, 2018). These three studies illustrate how generating multiple parallel omic datasets targeted toward signaling pathway and metabolism can yield very valuable insight to understand the molecular features of diseases. In the future, it is very likely that more multi-omic datasets will be generated to reconstruct a global regulatory picture of cellular functions. The methods discussed in 4. Footprint analysis and 5. Footprint based multi-omic network can be useful for the analysis of such multi-omic datasets. They can generate insights into cellular mechanisms spanning across signaling, regulatory and metabolic networks. Indeed, these methods mainly rely on principles that are conserved across signaling and metabolism, such as enzyme/substrate relationships, and are specifically designed to provide functional insights.

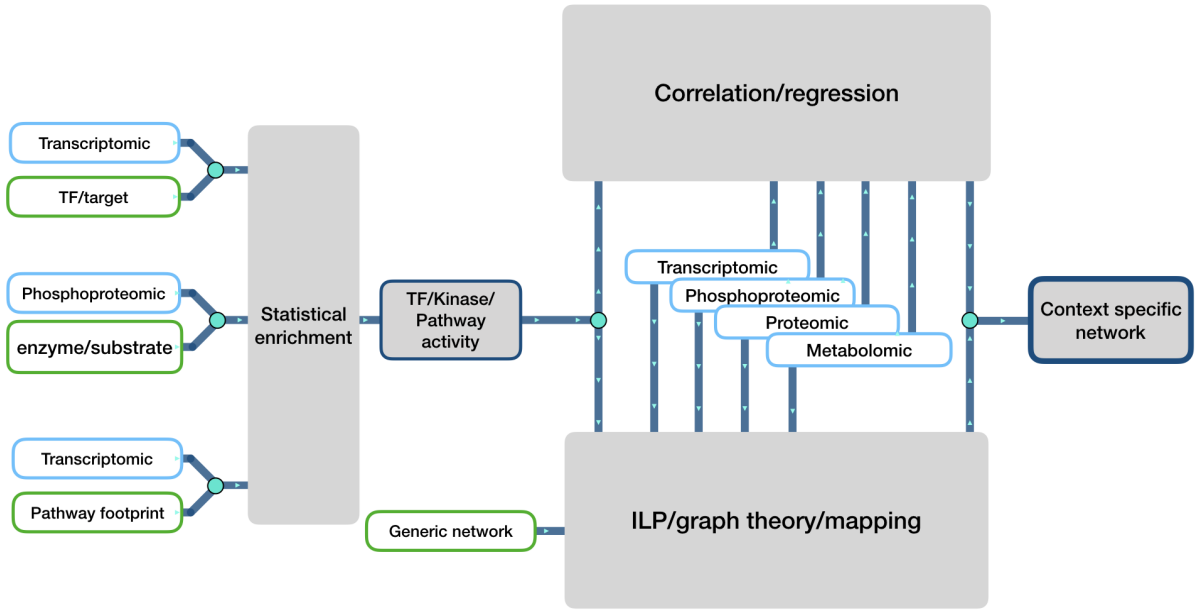


Figure 4 - Summarised representation of the multi-omic analysis workflow

On the left, statistical enrichment analysis is used to estimate activity of kinases, transcription factors and pathways. Then, multiple types of omic data can be connected together with these activities by correlation/regression methods. They can also be combined with prior knowledge networks through network contextualisation methods (optimisation, graph theory and mapping). Finally, the output of network contextualisation and correlation-based methods can be used, independently or combined, to generate multi-omic context specific networks.

Chapter 2 : Omic data exploratory and functional analysis in various contexts

During the course of my PhD, I developed omic data analysis skills (such as the ones presented in Chapter 1) that allowed me to be involved in a variety of collaboration with experimental laboratories that generated such datasets. While such collaborations were subject to a severe attrition from initial data analysis to actual publication, a handful of projects actually yielded significant publication. In this chapter, I present highlights from these projects and give some details about my involvement and contribution to each of them.

1. Kinetic modelling of quantitative proteome data predicts metabolic reprogramming of liver cancer(Berndt *et al*, 2020)

In this paper, I analysed a proteomic dataset generated from biopsies of liver tumor and healthy liver tissues. I first applied a naive exploratory analysis on all samples to get a feel of what the dataset looked like. This comprised hierarchical clustering of the proteomic abundance profiles and also of their corresponding cross-correlation matrix, and principal component analysis. Thus, I made sure that the proteomic dataset could clearly discriminate between the healthy and tumor samples. Then I performed a differential analysis using LIMMA(Ritchie *et al*, 2015), an R package that boosts the hypothesis testing statistical power by leveraging information shared across all the tested features. Finally, I performed a pathway enrichment analysis using the Piano R package. Piano estimates enrichment scores by integrating the scores of multiple enrichment analysis algorithms together (such as GSEA or PAGE). It also allows recovery of pathways that are significantly enriched with subsets of genes that are regulated in opposing directions. This supported the hypothesis that glycolysis was strongly deregulated in liver tumors. Overall, the proteomic data analysis results were coherent with the hypothesis generated by the kinetic modelling of tumor metabolism and helped to support the model's hypotheses.

2. NADH Shuttling Couples Cytosolic Reductive Carboxylation of Glutamine with Glycolysis in Cells with Mitochondrial Dysfunction(Gaude *et al*, 2018)

In this paper, we dissected the molecular consequences of mitochondrial dysfunction in tumor cell lines. Thus, I analysed a proteomic dataset generated from a new cell line model of mitochondrial dysfunctions. This cell line effectively allows the experimentalist to culture them with controlled levels of mitochondrial dysfunction. I applied the same analysis pipeline as the ones described in the previous part (Kinetic modelling...). This time, the analysis notably highlighted that cells were displaying levels of cytoskeletal and cell mobility deregulation that were proportional to the level of mitochondrial dysfunction.

3. Gli1+ Mesenchymal Stromal Cells Are a Key Driver of Bone Marrow Fibrosis and an Important Cellular Therapeutic Target(Schneider *et al*, 2017)

In this paper, we studied a subpopulation of stromal cells in bone marrow that is suspected to play a critical role in bone marrow fibrosis progression (Gli1+ stromal cells, or fibrosis driving stromal cells). I analysed a transcriptomic dataset generated from fibrosis driving stromal cells. The stromal cells were extracted from inducible bone marrow fibrosis mouse models, in healthy and bone marrow fibrosis conditions. I applied the same analysis pipeline as in the first study (Kinetic modelling...). The pathway enrichment analysis was able to highlight the strong deregulation of inflammation associated metabolic pathways such as leukotriene and prostaglandin pathways. Interestingly, both pathways use the same precursor, the arachidonic acid. This finding was particularly interesting as it was able to connect metabolic deregulation with a tissue-level phenotype (inflammation). The pathway enrichment analysis also allowed the analysis on the CXCL4 gene. This gene was particularly deregulated in bone marrow fibrosis and seemed to be a major driver of the inflammation and fibrosis progression. I also estimated the activity changes of pathways and transcription factors in bone marrow fibrosis using Progeny and DOROTHEA. While the pathway and TF activities were not directly reported in this paper, they helped shape up the follow up analysis that I present later in more detail (Increased CXCL4 expression...).

4. Increased CXCL4 expression in hematopoietic cells links inflammation and progression of bone marrow fibrosis in myeloproliferative neoplasms(Gleitz *et al*, 2020)

In this paper, we followed up on our previous work with fibrosis driving stromal cells (Gli1+ Mesenchymal Stromal Cells...) with a focus on the role of CXCL4. A cell co-culture model was established with megakariocytes and stromal cells extracted from fibrotic bone marrow. Transcriptomic datasets were generated from these cells at early and late time points of fibrosis progression. I applied the same pipeline of differential analysis and pathway enrichment analysis as presented previously (Kinetic modelling...). This showed that fibrosis driving stromal cells displayed dramatically different pathway activity profiles in early and late time points of fibrosis progression (Figure 1). Furthermore, this highlighted the fact that CXCL4 was actually over-expressed in megakariocytes but not in fibrosis driving stromal cells at an early fibrosis progression time point. CXCL4 seemed to be over-expressed in fibrosis driving stromal cells only at a late progression time point. This further supported the hypothesis that CXCL4 was actually not initially produced in fibrosis driving stromal cells, but rather they were aberrantly expressed in megakaryocytes at the start of the fibrotic transformation. CXCL4 seems to serve as a mediator for megakaryocytes to recruit and reprogram fibrosis driving stromal cells.

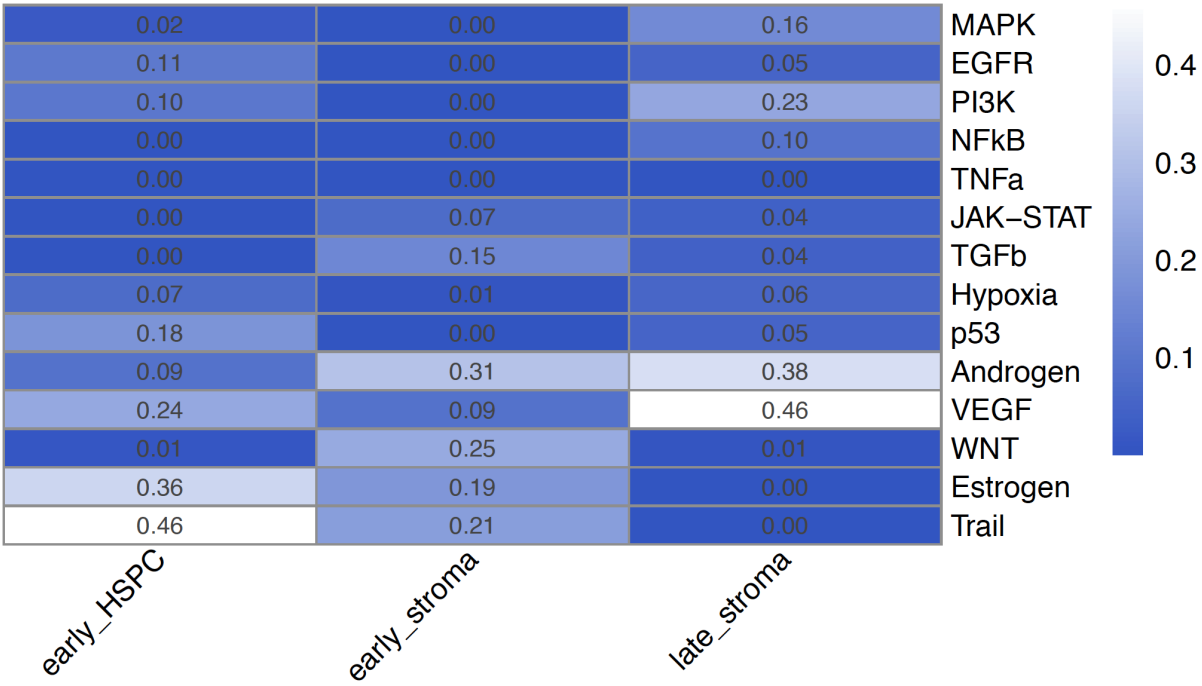


Figure 1 - Significance of progeny pathway activity changes in early and late fibrotic stromal cells and early fibrotic hematopoietic stem cells.

Then, a CXCL4 megakariocyte knockout experiment was designed to validate this hypothesis. Knocking out CXCL4 in megakaryocytes indeed partially recovered the fibrotic phenotype. I analysed a transcriptomic dataset generated from the co-culture CXCL4 KO model to understand better the molecular role of CXCL4 in the fibrosis progression. I compared the effect of fibrosis induction in megakaryocytes and fibrosis driving stromal cells both in WT and KO conditions. I used progeny and DOROTHEA to characterise pathway and TF activities in these different conditions. This highlighted that the JAK-STAT pathway activity was strongly down-regulated in stromal cells when fibrosis was induced in the CXCL4 knockout condition. This stands opposed to its up-regulation in stromal cells when fibrosis is induced in the WT condition. Overall, the CXCL4 KO markedly reduced the activity of pro-inflammatory pathways (Trail, NFkB and TNFalpha) in stromal cells when fibrosis was induced. As a next step, we hope to use the co-culture model to study more deeply the cell to cell communication by connecting TF and pathway deregulations across different cell types (Figure 2).

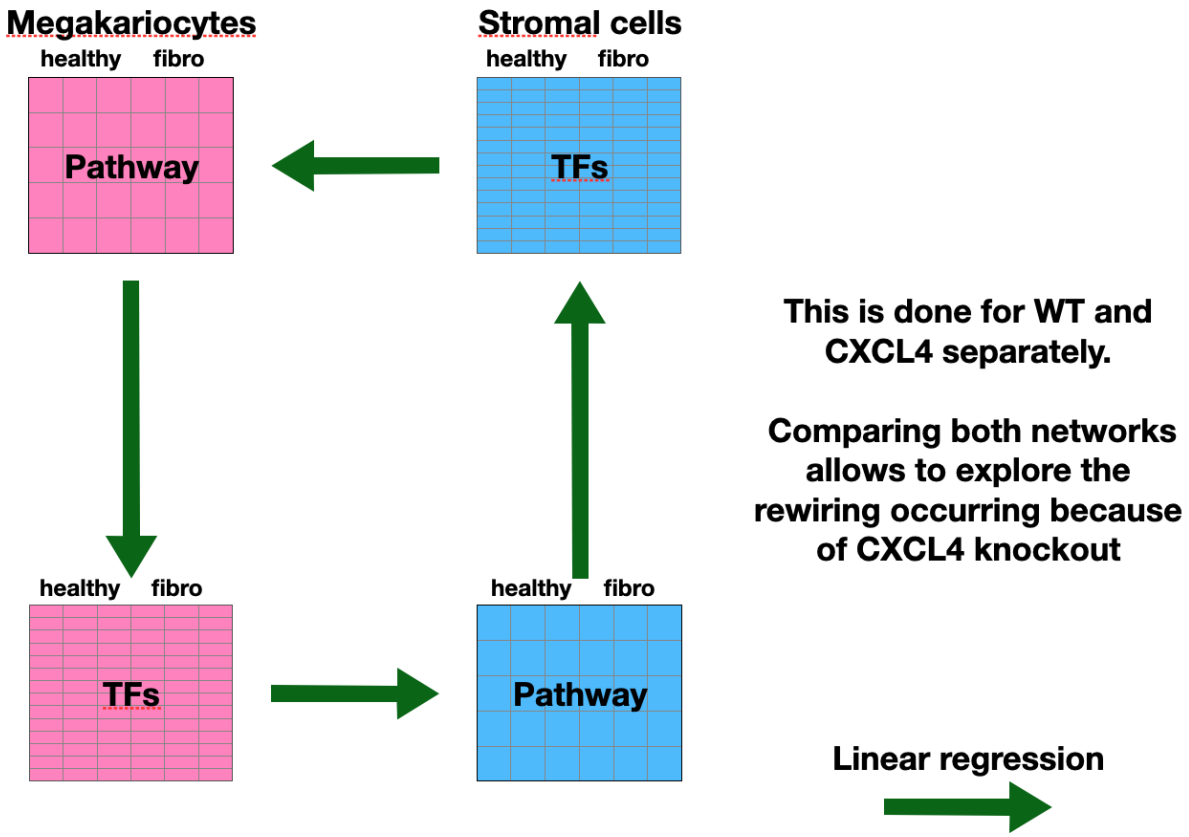


Figure 2 - Schematic representation of the first steps to study cell/cell communications between megakaryocytes and stromal cells.

We connect progeny and TF activity scores with simple linear regression. The linear models predict TF activities from pathway activities in the context of intracellular connections. The model direction is the opposite in the context of cell to cell communications, with pathway activities from one cell population being predicted from TF activities from another cell population.

5. Proteomes in 3D: in situ protein structural states as a readout for proteome functional alterations (Cappelletti *et al*, 2021)

In this paper, a new mass-spectrometry based method is presented. This method allows to measure the abundance changes of specific conformations of proteins at a large scale, allowing us to look at proteomic data from a completely new angle. A dataset of proteomic conformation changes was generated from yeast submitted to osmotic stress. In parallel, the experimentalist also generated a phosphoproteomic dataset from the same yeast culture. I estimated kinase/phosphatase activity changes from the phosphoproteomic dataset and I

systematically highlighted protein conformational changes that could be explained by changes in the activity of upstream kinases and phosphatases. I notably highlighted how SNF1, STE20, PBS2 and HOG1, canonical responder of osmotic stress, were displaying cascading changes of protein conformation and kinase activities (Figure 3).

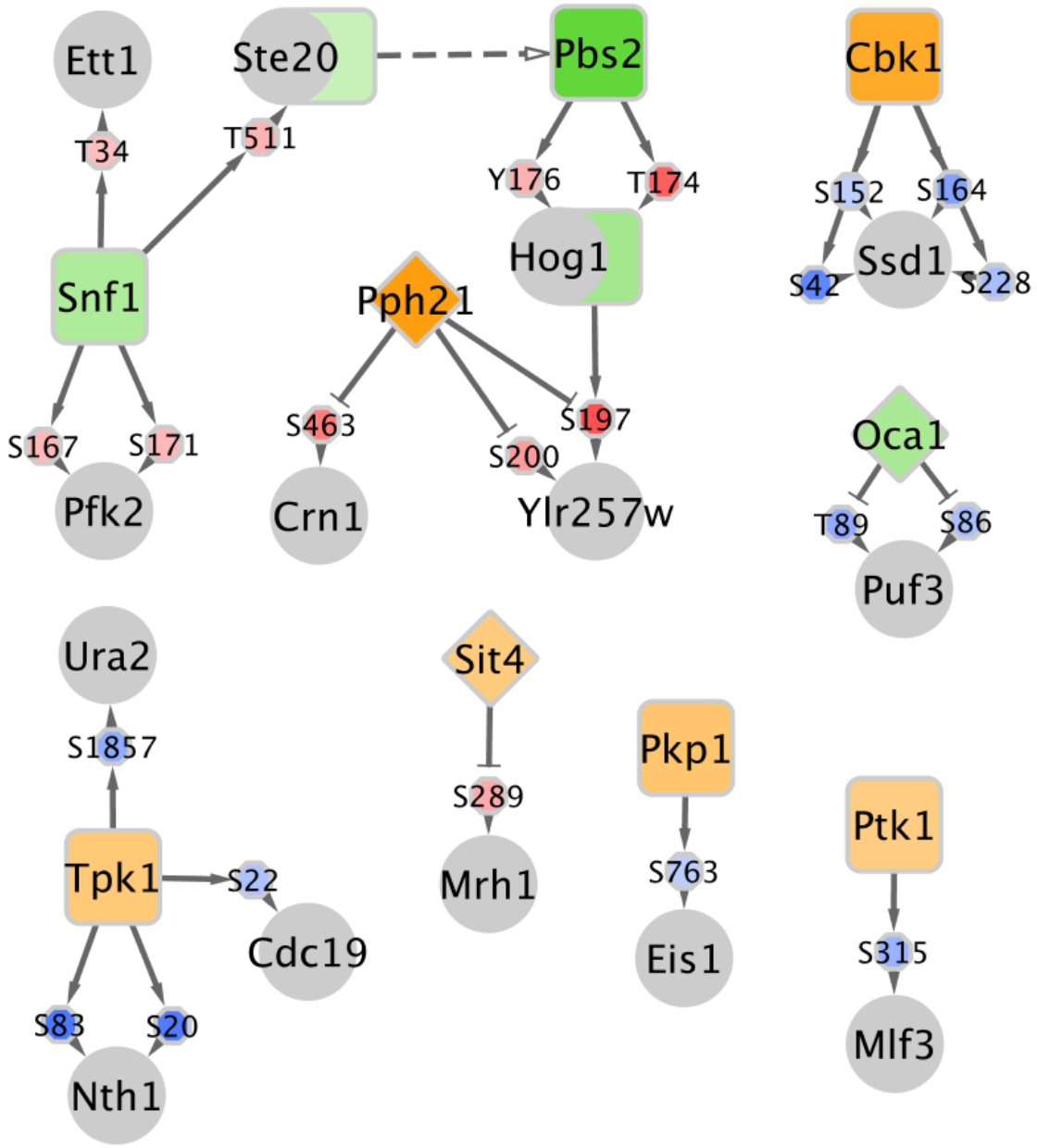


Figure 3 - Network representation of kinase/phosphatase activity changes with their target phosphorylation and conformational changes.

Square/diamond represent kinase/phosphatases. Small octogones represent phosphorylation sites. Grey circles are proteins that display at least one peptide with significant conformational change. Green/orange represent up/down regulation of kinase and phosphatase activities. Red/Blue represent

up/down-regulation of phosphorylation site abundance.

6. SREBP1-induced fatty acid synthesis depletes macrophages antioxidant defences to promote their alternative activation (Bidault *et al.* 2021)

In this paper, molecular determinants of macrophage activation are studied. It especially highlighted the link between metabolic reprogramming of fatty acid to support generation of Radical Oxidative Species and subsequent activation of macrophages. I notably analysed transcriptomic data from SCAP knockout macrophages exposed to IL4 (IL4 is a known activator of macrophages). This notably showed that oxidative stress response pathways were significantly down-regulated when SCAP was knocked out compared to wild-type macrophages. Since SCAP is a notable activator of SREBP1, which supported the hypothesis that SREBP1 was a critical intermediate of macrophage metabolic reprogramming to support their IL4 dependent activation.

7. The Global Phosphorylation Landscape of SARS-CoV-2 Infection (Bouhaddou *et al*, 2020)

In this paper, a functional profile of Sars-Cov-2 infection is built from a large phosphoproteomic dataset generated from airway derived cells infected by Sars-cov-2. This analysis helped to understand the signaling pathway reprogramming following infection by Sars-Cov-2 and to propose and validate in-vitro potential new therapeutic targets to block virus proliferation. I set up the part of the analysis pipeline that served as a template to analyse transcriptomic data generated from the same conditions. The analysis pipeline covered differential analysis, TF and pathway activity estimation, and subsequent signaling pathway contextualisation with CARNIVAL. The results of the TF enrichment analysis were put in perspective of kinase activity estimations and showed that the downstream transcription factors of the deregulated p38/MAPK pathway were coherently among the top deregulated TFs in airway derived cell lines infected by Sars-Cov-2.

Chapter 3 : Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses

Chapter 3 is a preliminary version of a manuscript that was later published in bioRxiv and Molecular Systems Biology : Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses (Dugourd *et al*, 2021). This work was also featured as the cover of Molecular Systems Biology (Volume 17; Issue 1). The text and figures used in this chapter were written solely by A. Dugourd.

Abstract

Multi-omics datasets can provide molecular insights beyond the sum of individual omics. Diverse tools have been recently developed to integrate such datasets, but there are limited strategies to systematically extract mechanistic hypotheses from them. Here, we present COSMOS (Causal Oriented Search of Multi-Omics Space), a method that integrates phosphoproteomics, transcriptomics, and metabolics datasets. COSMOS combines extensive prior knowledge of signaling, metabolic, and gene regulatory networks with computational methods to estimate activities of transcription factors and kinases as well as network-level causal reasoning. COSMOS provides mechanistic hypotheses for experimental observations across multi-omics datasets. We applied COSMOS to a dataset comprising transcriptomics, phosphoproteomics, and metabolomics data from healthy and cancerous tissue from nine renal cell carcinoma patients. We used COSMOS to generate novel hypotheses such as the impact of Androgen Receptor on nucleoside metabolism and the influence of the JAK-STAT pathway on propionyl coenzyme A production. We expect that our freely available method will be broadly useful to extract mechanistic insights from multi-omics studies.

1. Introduction

"Omics" technologies measure at the same time thousands of biological molecules in biological samples, from DNA, RNA and proteins to metabolites. Omics datasets are an essential component of systems biology, and are made possible by the popularization of analytical methods such as Next Generation Sequencing or Mass-Spectrometry. Omics data have enabled the unbiased characterization of the molecular features of multiple human diseases, particularly in cancer(Iorio *et al*, 2016; Jelinek & Wu, 2012; Subramanian *et al*, 2017). It is becoming increasingly common to characterize multiple omics layers in parallel, with so-called "trans-omics analysis", to gain biological insights spanning multiple types of cellular processes(Sciacovelli *et al*, 2016; Kawata *et al*, 2018; Vitrinel *et al*, 2019). Consequently, many tools are developed to analyze such data(Argelaguet *et al*, 2018; Sharifi-Noghabi *et al*, 2019; Tenenhaus *et al*, 2014; Singh *et al*, 2019; Liu *et al*, 2019b), mainly by adapting and combining existing "single omics" methodologies to multiple parallel datasets. These methods identify groups of measurements and derive integrated statistics to describe them, effectively reducing the dimensionality of the datasets. These methods are useful to provide a global view on the data, but additional processing is required to extract mechanistic insights from them.

To extract mechanistic insights from datasets, some methods (such as pathway enrichment analysis) use prior knowledge about the players of the process being investigated. For instance, differential changes in the expression of the genes that constitute a pathway gene expression are used to infer the activity of that pathway. Methods that a priori define groups of measurements based on known regulated targets (that we call footprints(Dugourd & Saez-Rodriguez, 2019b)) of transcription factors (TFs)(Alvarez *et al*, 2016; Garcia-Alonso *et al*, 2019), kinases/phosphatases(Wiredja *et al*, 2017) and pathway perturbations(Schubert *et al*, 2018), provide integrated statistics that can be interpreted as a proxy of the activity of a molecule or process. These methods seem to estimate more accurately the status of processes than classic pathway methods (Cantini *et al*, 2018; Dugourd & Saez-Rodriguez, 2019b; Schubert *et al*, 2018). Since each of these types of footprint methods work with a certain type of omics data, finding links between them could help to interpret them collectively in a mechanistic manner. For example, one can use a network diffusion algorithm, such as TieDIE(Paull *et al*, 2013), to connect different omics footprints together(Drake *et al*, 2016). This approach provides valuable insights, but diffusion (or random walk) based algorithms do not typically take into account causal information (such

as activation/inhibition) that is available and are important to extract mechanistic information. TieDIE partially addressed this problem by focusing the diffusion process on causally coherent subparts of a network of interest, but it is thus limited to local causality.

Recently, we proposed the CARNIVAL tool(Liu *et al*, 2019a) to systematically generate mechanistic hypotheses connecting TFs through global causal reasoning supported by Integer Linear Programming. CARNIVAL connects activity perturbed nodes such as drug targets with deregulated TFs activities by contextualizing a signed and directed Prior Knowledge Network (PKN). We had hypothesized how such a method could potentially be used to actually connect footprint based activity estimates across multiple omics layers(Dugourd & Saez-Rodriguez, 2019b).

In this study, we introduce COSMOS (Causal Oriented Search of Multi-Omics Space), an approach that builds on CARNIVAL to connect TF and kinase/phosphatases activities as well as metabolite abundances with a novel PKN spanning across multiple omics layers (Figure 1). COSMOS uses CARNIVAL's Integer Linear Programming (ILP) optimization strategy to find the smallest coherent subnetwork causally connecting as many deregulated TFs, kinases/phosphatases and metabolites as possible. The subnetwork is extracted from a novel integrated PKN spanning signaling, transcriptional regulation and metabolism of > 67000 edges. CARNIVAL's ILP formulation effectively allows to evaluate the entire network's causal coherence given a set of known TF, kinases/phosphatases activities and metabolite abundances. While we showcase this method using transcriptomics, phosphoproteomics and metabolomics inputs, COSMOS can theoretically be used with any other additional inputs, as long as they can be linked to functional insights (for example, a set of deleterious mutations). As a case study, we generated transcriptomics, phosphoproteomics, and metabolomics datasets from kidney tumor tissue and corresponding healthy kidney tissue out of nine clear cell renal cell carcinoma (ccRCC) patients. We estimated changes of activities of TFs and kinase/phosphatases as well as metabolite abundance differences between tumor and healthy tissue. We integrated multiple curated resources of interactions between proteins, transcripts and metabolites together to build a trans-omics PKN. Next, we contextualized the trans-omics PKN to a specific experiment. To do so, we identified causal pathways from our prior knowledge that connect the observed changes in activities of TFs, kinases, phosphatases and metabolite abundances between tumor and healthy tissue. These causal pathways can be used as hypothesis generation tools to better understand the molecular phenotype of kidney cancer.

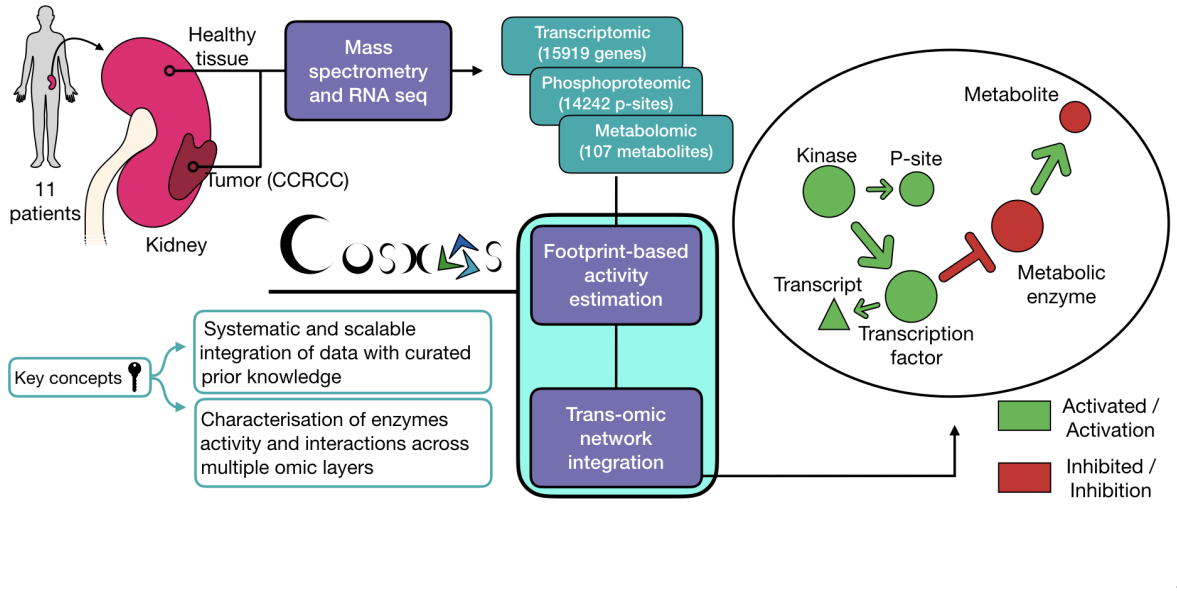


Figure 1 - Overview of analysis pipeline

From left to right: We sampled and processed 11 patient tumors and healthy kidney tissues from the same kidney through RNA sequencing and 9 of those same patients through mass-spectrometry to characterise their transcriptomics, phospho-proteomics, and metabolomics profiles. We calculated differential abundance for each detected gene, phospho-peptide and metabolite. We estimated kinase and transcription factor activities using the differential analysis statistics and footprint-based methods. We used the estimated activities alongside the differential metabolite abundances to contextualise (i.e. extract the subnetwork that better explains the phenotype of interest) a generic trans-omics causal network.

2. Results

2.1 Building the multi-omics dataset

To build a multi-omics dataset of renal cancer, we performed transcriptomics, phosphoproteomics, and metabolomics analyses of renal nephrectomies and adjacent normal tissues of renal cancer patients (for details on the patients see methods). First, we processed the different omics datasets to prepare for the analysis. For the transcriptomics dataset, 15919 transcripts with average counts > 50 were kept for subsequent analysis. In the phosphoproteomics dataset, 14243 phosphosites detected in at least four samples were kept. In the metabolomics dataset 107 metabolomics detected across 16 samples were kept. Principal Component Analysis (PCA) of each omics dataset independently showed a clear separation of healthy and tumor tissues on the first component (transcriptomics : 40% of

explained variance (EV), phosphoproteomics : 26% of EV, metabolomics : 28% of EV, Supplementary Figure 1), suggesting that tumor sample displayed molecular deregulations spanning across signaling, transcription and metabolism. Each omics dataset was independently submitted to differential (tumor vs healthy tissue) analysis using LIMMA(Ritchie *et al*, 2015). We obtained 6699 transcript and 21 metabolites significantly regulated with False Discovery Rate (FDR) < 0.05. While only 11 phosphosites were found under 0.05 FDR, 447 phosphosites had an FDR < 0.2. This result confirmed that tumor samples displayed molecular deregulations spanning across signaling, transcription, and metabolism but that TF dysregulation is more pervasive. The differential statistics for all transcripts, phospho-proteins and metabolites were then used for further downstream analysis.

2.2 Footprint based transcription factor, kinase and phosphatase activity estimation

We then performed computational footprint analysis to estimate the activity of proteins responsible for changes observed in specific omics datasets. For transcriptomics and phosphoproteomics data, this analysis estimates transcription factor and kinases/phosphatase activity, respectively. 32586 Transcription Factor (TF) to target interactions (i. e. transcript under the direct regulation of a transcription factor) were obtained from DOROTHEA(Garcia-Alonso *et al*, 2019), a meta-resource of TF-target interactions. Those TF-target interactions span over 452 unique transcription factors. In parallel, 33616 interactions of kinase/phosphosphate and their phosphosite targets (i. e. phosphopeptides directly (de)phosphorylated by specific kinases(phosphatases)) were obtained from Omnipath(Türei *et al*, 2016) kinase substrate network, a meta resource focused on curated information on signaling processes. Only TFs and kinases/phosphatases with at least 25 and 5 detected substrates, respectively, were included. This led to the activity estimation of 229 TFs and 174 kinases. In line with the results of the differential analysis, where fewer phosphosites were deregulated than transcripts, TF activities displayed a stronger deregulation than kinases. TF activity scores reached a maximum of eight standard deviations (sd) for Transcription Factor AP-2 Gamma (TFAP2C) (compared to the null score distribution) while kinase activity scores reached a maximum of 4.6 sd for Casein Kinase 2 Alpha 1 (CSNK2A1). In total, 102 TFs and kinases/phosphatase had an absolute score over 1.7 sd (p-val<0.05) and were considered significantly deregulated in kidney tumor samples. The presence of several known signatures of ccRCC corroborated the validity of our analysis. For instance, hypoxia (HIF1A, EPAS1), inflammation (STAT1/2) and oncogenic

(MYC, Cyclin Dependent Kinase 2 and 7 (CDK2/7)) markers were up-regulated in tumors compared to healthy tissues (Figure 2). Furthermore, among suppressed TFs we identified, HNF4A has been previously associated with ccRCC(Lucas *et al*, 2005).

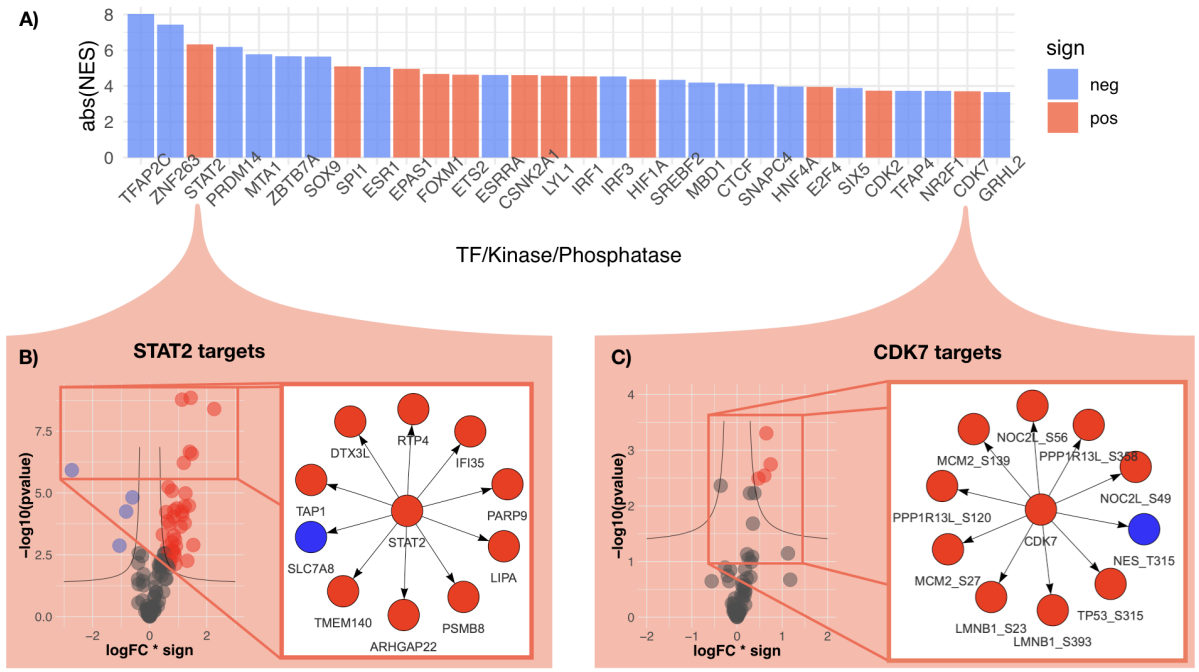


Figure 2 TF, kinase and phosphatase activities that change the most between cancer and healthy tissue

A) Bar plot displaying the Normalised Enrichment Score (NES, proxy of activity change) of the 30 most changing TF, kinase and phosphatases activities between kidney tumor and adjacent healthy tissue. Blue/red color represent the sign of the activity change (negative/positive, respectively). B) Right panel shows the 10 most changing RNA abundances of the STAT2 regulated transcripts. Left panel shows the change of abundances of all STAT2 regulated transcripts that were used to estimate its activity change. X axis represents log fold change of regulated transcripts multiplied by the sign of regulation (-1 for inhibition and 1 for activation of transcription). Y axis represents the significance of the log fold change (-log10 of p-value). C) Right panel shows the 10 most changing phospho-peptide abundances of the CDK7 regulated phospho-peptides. Left panel shows the change of abundances of all CDK7 regulated phospho-peptides that were used to estimate its activity change.

2.3 Causal network analysis

We set out to find potential causal mechanistic pathways that could explain the changes we observed in TF, kinases/phosphatase activities, and metabolic abundances. Thus, we developed a systematic approach to search in public databases, via OmniPath, for plausible causal links between significantly deregulated TFs, kinases/phosphatases and metabolites. In brief, we investigated if changes in TF, kinase/phosphatase activities, and metabolite

abundance can explain each other with the support of literature-curated molecular interactions. An example of such a mechanism can be the activation of the transcription of MYC gene by STAT1. Since both STAT1 and MYC display increased activities in tumors, and there is evidence in the literature that STAT1 can regulate MYC transcription(Kharma *et al*, 2014; Ramana *et al*, 2000), it may indicate that this mechanism is responsible for this observation.

First, we needed to map the deregulated TFs, kinases and metabolites on a causal prior knowledge network spanning over signaling pathways, gene regulation, and metabolic networks. Hence, we combined multiple sources of experimentally curated causal links together to build a trans-omics causal prior knowledge network (trans-omics PKN). This trans-omics PKN must include direct causal links between proteins (kinase to kinase, TF to kinase, TF to metabolic enzymes, etc…), between proteins and metabolites (reactants to metabolic enzymes and metabolic enzymes to products) and between metabolites and proteins (allosteric regulations). High confidence (>= 900 combined score) allosteric regulations of the STITCH database(Szklarczyk *et al*, 2016) were used as the source of causal links between metabolites and enzymes (Figure 3A). The directed signed interactions of the Omnipath database were used as a source of causal links between proteins (Figure 3B). The human metabolic network Recon3D(Brunk *et al*, 2018) (without cofactors and hyper-promiscuous metabolites, see methods) was converted to a causal network and used as the source of causal links between metabolites and metabolic enzymes (Figure 3C). The resulting trans-omics PKN consists of 69517 interactions and contains causal paths linking TF/kinase/phosphatase with metabolites and vice-versa in a machine readable format. This network is available at http://metapkn.omnipathdb.org/.

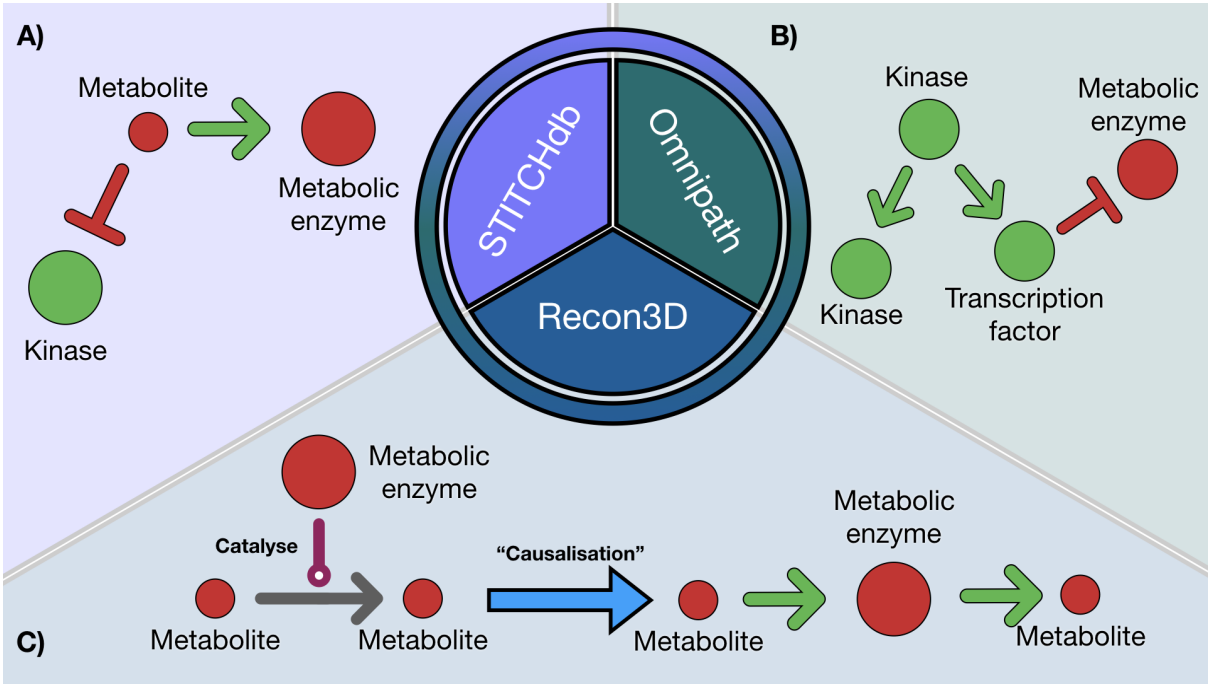


Figure 3 - Graphical explanation of trans-omics PKN sources

Schematic representation of the trans-omics generic network (trans-omics PKN) created combining STITCHdb, Omnipath and Recon3D. A) STITCHdb provides information on inhibition/activation of enzyme activities mediated by metabolites. B) Omnipath provides information inhibition/activation of enzyme activities mediated by other enzymes based mainly on curated resources. C) Recon3D provides information on reactants and products associated with metabolic enzymes. To make this information compatible with the causal edges from Omnipath and STITCH, the interactions of recon3D are converted so that reactants "activate" their metabolic enzymes, which themselves "activate" their products.

We then used the trans-omics PKN to systematically search causal paths between the deregulated TFs, kinases/phosphatases and metabolites. The CARNIVAL(Liu *et al*, 2019a) tool uses Integer Linear Programming (ILP) to find causal paths between perturbations and deregulated TFs using a PKN and infers the state of intermediate nodes when it is unknown. Here we use CARNIVAL with our trans-omics PKN to find the smallest sign-coherent subnetwork connecting as many deregulated TFs, kinases/phosphatases, and metabolites as possible. CARNIVAL is first used to find causal paths going from TFs/kinases/phosphatases to the metabolites (the 'forward network'). Then, in order to complete the loop, CARNIVAL is used to go from metabolites to TFs/Kinases/phosphatases ('backward network').

When applied to our kidney cancer data, the two resulting (forward and backward) networks are then combined into a single network of 250 signed directed interactions (Supplementary Figure 2). These interactions are directly interpretable as mechanistic hypotheses. We present some of them using official symbol nomenclature for genes and metabolites. For example, it appears that Androgen Receptor (AR) activity inhibition could be responsible for the observed downregulation of uridine, adenine, and inosine metabolism by down-regulating the expression of ACPP, DBI, and SMS metabolic enzymes (Figure 4A). Of note, AR expression has a protective role in ccRCC progression (Zhao *et al*, 2016; Zhu *et al*, 2014). Interestingly, the COSMOS network shows adenine depletion could lead to adenosine depletion (since adenosine can be produced from adenine). Adenosine is a known activator of the C-X-C Motif Chemokine Receptor 4 (CXCR4) (Rolland-Turner *et al*, 2013; Richard *et al*, 2006), so its depletion could lead to the predicted down-regulation of CXCR4 activity. The combined AR and CXCR4 down-regulation might indicate that these tumors are not metastatic (Wang *et al*, 2017; Vanharanta *et al*, 2013; Rodrigues *et al*, 2018). The COSMOS network also shows that CXCR4 regulates Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Gamma (PIK3CG), which itself regulates 3-Phosphoinositide Dependent Protein Kinase 1 (PDPK1). Thus, CXCR4 down-regulation could then explain PDPK1 activity down-regulation through the inhibition of PIK3CG. COSMOS further proposes that the activation of JAK kinase would be a good explanation for the apparent activation of STAT transcription factors in the tumor, leading to activation of IRF1 and MYC (Figure 4B). Interestingly, JAK2 was found to be amplified in ccRCC (Network & The Cancer Genome Atlas Research Network, 2013). The STAT3 activation could explain the depletion of o-propanoylcarnitine due to the downregulation of metabolic enzymes responsible for the transport of its precursor, Sterol Carrier Protein 2 (SCP2). CDK2 could itself explain the activity of ATM and TP53 through Forkhead Box M1 (FOXM1) and Aurora Kinase B (AURKB) signaling, leading to the activation of Dual specificity tyrosine-phosphorylation-regulated kinase 2 (DYRK2), HIF1A and the accumulation of L-Glutamine (Figure 4C). FOXM1 was recently highlighted as a particularly important driver of metabolic changes in ccRCC(Pandey *et al*, 2020). Finally, the COSMOS network shows that the down-regulation of PDPK1 appears as a good explanation for L-Citrulline accumulation and ethanolamine depletion, by indirectly modulating the activity of Nitric Oxide Synthase 1 (NOS1) and Phospholipase D1 (PLD1) metabolic enzymes (Figure 4D). Furthermore, PDPK1 directly controls the activity of the Protein Kinase C protein family (PRKCA, PRKCD, PRKCE and PRKACA). These kinases are known to be involved in metastasis progression (Brenner *et al*, 2003; Engers *et al*, 2000). Their down-regulation

predicted by COSMOS further supports the idea that these tumors are not metastatic. These results demonstrate how the pipeline can be used to extract relevant mechanistic hypotheses explaining the enzymatic and metabolic deregulations at signaling and transcriptional levels.

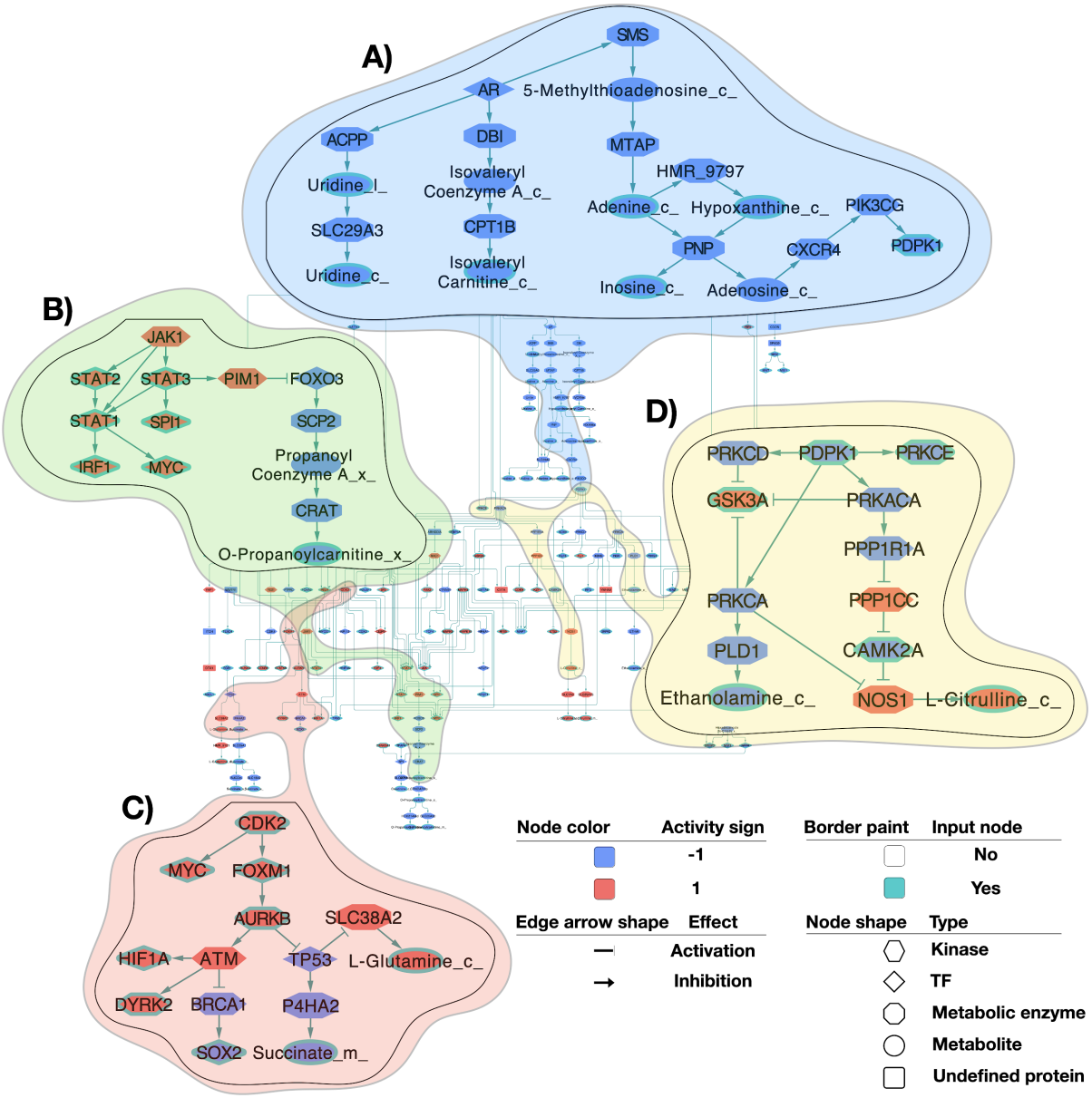


Figure 4 - Systematically generated mechanistic hypotheses explaining changing TF, kinase, phosphatase activities and metabolic abundances

COSMOS generates mechanistic hypotheses which are represented in the form of a context specific causal network. This network links the significant changes in estimated enzyme activities and metabolic abundance (192 nodes and 250 edges). Diamond shapes represent TFs , hexagon shape represents kinases and phosphatases, octagon shape represents metabolic enzymes and ellipse shape represents metabolites. Blue/red color represents inhibited/activated enzyme activities and

depleted/accumulated metabolites. Edges with arrowheads represent activatory interactions and T-shaped ones represent inhibitory interactions. A full-page version of this figure is available as Supplementary Figure 2 A) ,B) ,C) and D) represent subnetworks extracted to zoom on specific hypotheses. For example, B) represents how AR activity can lead to the inhibition of nucleotides and isovaleryl carnitine synthesis observed in tumors compared to healthy tissues, in turn explaining the inhibited activity of PDPK1.

2.4 Consistency analysis

Due to the combined effect of experimental noise and incompleteness of prior knowledge (kinase/substrate interactions, TF/targets interactions and meta PKN), it is critical to assess the performance of the pipeline presented above. We first looked if some of the generated hypotheses (see 2.3) were supported by parts of the datasets that were not directly used by CARNIVAL (Supplementary Figure 3). We couldn't estimate the True Negative Rate of CARNIVAL in this multi-omics context. Indeed, nodes that are not integrated in the final subnetwork by CARNIVAL are simply not considered informative to explain the relationship between the input protein activities and metabolites. Yet, that doesn't inform us on their actual functional state. Consequently, we focused on the True Positive Rate (TPR), for which we had reasonable estimates. The TF activity displayed by CARNIVAL can come from two distinct sources. The first source consists of the original footprint based activity estimation (using DOROTHEA and transcript abundances of target genes). The second source consists of actual CARNIVAL activity predictions based on molecular signal propagation through activating/inhibiting links (not using transcript abundances) connecting TF, kinases and phosphatases together. This is the case as some TFs can serve as intermediate links to connect upstream perturbations with downstream nodes. Consequently, in the case of a TF, CARNIVAL will implicitly model its action on the direct downstream targets. Thus, for every TF/target regulation of the CARNIVAL network, we checked whether the change of abundance of the target transcripts was actually coherent with the predicted activity of the TF displayed by CARNIVAL. We tested this over a range of differential transcript abundance t-value threshold between 0 and 2. Nine transcripts were regulated by TF whose activity was predicted by CARNIVAL only (second source), that is, the transcripts were not used as inputs to build the COSMOS model. Out of those nine transcripts, the TPR ranged between 0.62 and 0.15 depending on the t-value threshold (n = 13) (Supplementary Figure 4). It performed better than a random baseline for considered t-value thresholds ranging from 0 to 1.7.

Another way to estimate the performance is to check if the CARNIVAL mechanistic hypotheses correspond to correlations observed in tumor tissues. Thus, on the one hand, a topological driven coregulation network was generated from the CARNIVAL network. The assumption behind this network is that direct downstream targets of the same enzymes should be co-regulated. On the other hand, a data driven correlation network of TFs, kinases and phosphatases was generated from tumor tissues alone. Assuming thresholds of absolute values of correlation ranging between 0 and 1 to define true positive co-regulations, the comparison between the topological driven coregulation network and the data driven correlation network yielded a TPR of ranging between 0.6 and 0 (n = 157) for the carnival predictions (Supplementary Figure 4). It performed consistently better than a random baseline over the considered range of correlation coefficient thresholds. These results indicate that while some of the causal links predicted by CARNIVAL are potentially valid, some of them don't find direct support in the data at hand. Thus, we sought to investigate if some of the mechanistic hypotheses could be experimentally validated.

3. Discussion

In this paper, we present COSMOS, an analysis pipeline to systematically generate mechanistic hypotheses by integrating multi-omics datasets with a broad range of curated resources of interactions between protein, transcripts and metabolites. We have first shown how TF, kinase and phosphatase activities could be coherently estimated from transcriptomics and phosphoproteomics datasets using footprint based analysis. This is a critical step before further mechanistic exploration. Indeed, transcript and phosphosite usually offer limited functional insights by themselves as their relationship with corresponding protein activity is usually not well characterised. Yet, they can provide information on the activity of the upstream proteins regulating their abundances. Thus, the functional state of kinases, phosphatases, and TFs is estimated from the observed abundance change of their known targets, i. e. their molecular footprint. Thanks to this approach, we could simultaneously characterise protein functional states in tumors at the level of signaling pathway and transcriptional regulation. Key actors of hypoxia response, inflammation pathway and oncogenic genes were found to have especially strong alteration of their functional states, such as HIF1A, EPAS1, STAT1/2, MYC and CDK2. Loss of VHL is a hallmark of ccRCC, and is directly linked to the stability of the HIF (HIF1A and EPAS1) proteins found deregulated by our analysis(Maxwell *et al*, 1999; Ivan *et al*, 2001; Jaakkola *et al*, 2001). Finding these established signatures of ccRCC to be deregulated in our analysis is a confirmation of the validity of this approach.

We then used CARNIVAL with a novel trans-omics causal Prior Knowledge Network spanning signaling, transcription and metabolism to systematically find potential mechanisms linking deregulated protein activities and metabolite concentrations. To the best of our knowledge, this is the first attempt to integrate these three omics layers together in a systematic manner using causal reasoning. Previous methods studying signaling pathways with multi-omics quantitative datasets (Drake *et al*, 2016) connected TFs with kinases and they were limited by the preselected locally coherent subnetwork of the TieDIE algorithm. Introducing global causality with CARNIVAL along with metabolomics data allows us to obtain a direct mechanistic interpretation of links between proteins at different regulatory levels and metabolites. The goal of our approach is to find a coherent set of such mechanisms connecting as many of the observed deregulated protein activities and metabolite concentrations as possible. Using CARNIVAL is particularly interesting as all the proposed mechanisms between pairs of molecules (proteins and metabolites) have to be plausible not only in the context of their own pairwise interaction but also with respect to all other molecules that we wish to include in the model. For example, the proposed activation of MYC by STAT1 is further supported by IRF1 activation, because STAT1 is also known to activate IRF1. CARNIVAL allows us to scale this type of reasoning up to the entire PKN with all significantly deregulated protein activities and metabolites.

With our dataset, the resulting network showed that AR inhibition, a known tumor suppressor in kidney cancer (Uhlen *et al*, 2017), would be a good candidate to explain the inhibition of nucleotide metabolism. It also predicted a depletion of adenine and consequently the down-regulation of PDPK1 activity through CXCR4 (Figure 4A). Footprint analysis showed a down-regulation of PDPK1 (that is, the abundance of phosphorylation on its direct target phosphosites is decreasing) activity, which is surprising since its expression is usually associated with slower proliferation of kidney tumor cells(Zhou *et al*, 2019; Emmanouilidi & Falasca, 2017). Yet, the observed coordinated depletion of adenine, hypoxanthine and inosine strongly support the estimated down-regulation of PDPK1 activity. A consequence of PDPK1 activity down-regulation could also be the up-regulation of citrulline production by NOS1 (Figure 4C). COSMOS additionally predicted how JAK-STAT pathway activation could lead to an inhibition of the production of propanoyl-carnitine (Figure 4B). Diminution of carnitine and its derivative have been indeed previously observed in kidney cancer as a consequence of cachexia(Sayed-Ahmed, 2010). Finally we could show the importance of CDK2 as a master regulator of many kinases and transcription factors such as MYC,

AURKB, E2F4 and consequently TP53 and ATM activities (Figure 4D). In particular, AURKB, which directly controls TP53 and ATM activities, appears to be a promising marker of kidney cancer(Wan *et al*, 2019; Tang *et al*, 2017; Bertran-Alamillo *et al*, 2019).

Then we assessed the performances of the approach in two ways. First, we used some of the data that was not directly used by CARNIVAL (i. e. genes that were not used for TF activity estimation and correlation between TF/kinase/phosphatase activities) to check the coherence of CARNIVAL predictions. Second, we used a tumor specific correlation network of TF and kinase activities to compare it to the co-regulation predicted by CARNIVAL. This yielded encouraging results, though imperfect, underscoring the fact that the mechanisms proposed by COSMOS - like those by any similar tool - are hypotheses.

There are three main known limits to the predictions of COSMOS. First, the input data is incomplete. Only a limited fraction of all potential phosphosites and metabolites are detected by mass spectrometry. This means that we have no information on a significant part of the PKN; part of the unmeasured network is kept in the analyses and the values are estimated as intermediate ‘hidden values’. Second, not all regulatory events between TFs, kinase and phosphatases and their targets are known, and activity estimation is based only on the known regulatory relationships. Thus, many TFs, kinase and phosphatases are not included because they have no curated regulatory interactions or no detected substrates in the data. Third, and conversely, COSMOS will find putative explanations within the existing prior knowledge that may not be the true mechanism, in particular if the latter is not captured in our knowledge.

These problems mainly originate from the importance that is given to prior knowledge in this method. Since prior knowledge is never perfect, the next steps of improvement could consist in finding ways to extract more knowledge from the observed data to weight in the contribution of prior knowledge. For instance, one could use the correlations between transcripts, phosphosites and metabolites to quantify the interactions available in databases such as Omnipath. Importantly, any other omics that relate to active molecules (such as miRNAs or metabolic enzyme fluxes) or can be used to estimate protein activities through footprint approaches (such as DNA accessibility or PTMs other than phosphorylation) can be seamlessly integrated. Moreover, COSMOS was designed to work with bulk omics datasets, and it will be very exciting to find ways of applying this approach to single cell datasets. Encouragingly, the footprint methods that bring data into COSMOS seem fairly robust to the

characteristics of single-cell RNA data such as dropouts(Holland *et al*, 2020). Finally, we expect that in the future data generation technologies will increase coverage and our prior knowledge will become more complete, reducing the mentioned limitations. In the meantime, we believe that COSMOS is already a useful tool to extract causal mechanistic insights from multi-omics studies.

4. Methods

4.1. Sample collection and processing

We included a total of 22 samples from 11 renal cancer patients (6 men, age 65.0+/-14.31, 5 women, age 65.2+/-9.257(mean+/-SD)) for transcriptomics and a subset of 18 samples from 9 of these patients (6 men, age 65+/-14.31; 3 women, age 63.33+/-11.06(mean+/-SD)) for metabolomics and phosphoproteomics analysis. Patients underwent nephrectomy due to renal cancer. We processed tissue from within the cancer and a distant unaffected area of the same kidney.

For details about the sample processing to generate the omic data, see : https://www.biorxiv.org/content/10.1101/2020.04.23.057893v1

4.2 Data normalisation and differential analysis

In the phosphoproteomics dataset, 19285 unique phosphosites were detected across 18 samples. Visual inspection of the raw data PCA first 2 components indicated two major batches of samples. Thus, each batch was first normalised using the VSN R package(Välikangas *et al*, 2018; Huber *et al*, 2002). We removed p-sites that were detected in less than 4 samples, leaving 14243 unique p-site to analyse. Visual inspection of the PCA first two components of the normalised data revealed that the first batch of samples could itself be separated in 3 batches (4 batches across all samples). Thus, we used the removeBatchEffect function of LIMMA to remove the linear effect of the 4 batches. Differential analysis was performed using the standard sequence of lmFit, contrasts.fit and eBayes functions of LIMMA, with FDR correction.

For the transcriptomics data, counts were extracted from fast.q files using the RsubRead R package and GRCh37 (hg19) reference genome. Technical replicates were averaged, and genes with average counts under 50 across samples were excluded, leaving 15919 genes measured across 22 samples. In order to allow for logarithmic transformation, 0 count values were scaled up to 0.5 (similar to the voom function of LIMMA). Counts were then normalised

using the VSN R package function and differential analysis was performed with LIMMA package, in the same way as the phosphoproteomics data.
For the metabolomics data, 107 metabolites were detected in 16 samples. Intensities were normalised using the VSN package. Differential analysis was done using limma in the same manner as for phosphoproteomics and transcriptomics. All data is available at https://github.com/saezlab/COSMOS.

4.3 Footprint based analysis

TF-target collection was obtained from DOROTHEA A,B and C interaction confidence levels through the Omnipath webservice using the URL "http://omnipathdb.org/interactions?datasets=tfregulons&tfregulons_levels=A,B,C&genesymbols=1&fields=sources,tfregulons_level" (version of 2020 Feb 05). For the enrichment analysis, the viper algorithm(Alvarez *et al*, 2016) was used with the limma moderated t-value as gene level statistic(Zyla *et al*, 2017). The eset.filter parameter was set to FALSE. Only TFs with at least 25 measured transcripts were included.

Kinase-substrate collection was obtained using the default resource collection of Omnipath, with the URL "http://omnipathdb.org/ptms?fields=sources,references&genesymbols=1" (version of 2020 Feb 05). For the enrichment analysis, the viper algorithm was used with the limma limma moderated t-value as phosphosite level statistic. The eset.filter parameter was set to FALSE. Only TFs with at least 5 measured transcripts were included. All data is available at https://github.com/saezlab/COSMOS.

4.4 Meta PKN construction

In order to propose mechanistic hypotheses spanning through signaling, transcription and metabolic reaction networks, multiple types of interactions have to be combined together in a single network. Thus, we built a meta Prior Knowledge Network (PKN) from three online resources, to incorporate three main types of interactions. The three types of interactions are protein-protein interactions, metabolite-protein allosteric interactions and metabolite-protein interactions in the context of a metabolic reaction network. Protein-protein interaction were imported from omnipath with the URL http://omnipathdb.org/interactions?genesymbols=1 (version of 2019 Feb 05), and only signed directed interactions were included (is_stimulation or is_inhibition columns equal to 1). Metabolic-protein allosteric interactions were imported from the STITCH database (version of 2019 November 06), with combined confidence score >= 900 after exclusion of interactions relying mainly on text mining.

For metabolic-protein interactions in the context of metabolic reaction network, Recon3D was downloaded from https://www.vmh.life/#downloadview (version of 2019 Feb 19). Then, the gene rules (“AND” and “OR”) of the metabolic reaction network were used to associate reactants and products with the corresponding enzymes of each reaction. When multiple enzymes were associated with a reaction with an “AND” rule, they were combined together as a single entity representing an enzymatic complexe. Then, reactants were connected to corresponding enzymatic complexes or enzymes by writing them as rows of Simple Interaction Format (SIF) table of the following form : reactant;1;enzyme. In a similar manner, products were connected to corresponding enzymatic complexes or enzymes by writing them as rows of a Simple Interaction Format (SIF) table of the following form : enzyme;1;product. Thus, each row of the SIF table represents either an activation of the enzyme by the reactant (i.e. the necessity of the presence of the reactant for the enzyme to catalyse it’s reaction) or an activation of the product by an enzyme (e.i. the product presence is dependent on the activity of its corresponding enzyme). Most metabolite-protein interactions in metabolic reaction networks are not exclusive, thus measures have to be taken in order to preserve the coherence of the reaction network when converted to the SIF format. First, metabolites that are identified as “Coenzymes” in the Medical Subject Heading Classification (as referenced in the Pubchem online database) were excluded. Then, we looked at the number of connections of each metabolite and searched the minimum interaction number threshold that would avoid excluding main central carbon metabolites. Glutamic acid has 338 interactions in our Recon3D SIF network and is the most connected central carbon metabolite, thus any metabolites that had more than 338 interactions was excluded. An extensive list of Recon3D metabolites (pubchem CID) with their corresponding number of connections is available in supplementary table 2. Metabolic enzymes catalyzing multiple reactions were uniquely identified for each reaction to avoid cross-links between reactants and products of different reactions. Finally, exchange reactions were further uniquely identified according to the relevant exchanged metabolites, as to avoid confusion between transformation of metabolites and simply exchanging them between compartments. Finally, each network (protein-protein, allosteric metabolite-protein and reaction network metabolite-protein) was combined into a single SIF table. This network is available at http://metapkn.omnipathdb.org/.

4.5 Meta PKN contextualisation

Given a set of nodes with corresponding activities (-1, 0 or 1) and a causal PKN, CARNIVAL finds the smallest coherent signed subnetworks connecting as many of the given nodes as

possible. CARNIVAL needs a set of starting and end nodes to look for paths in between. TFs, kinases and phosphatases absolute normalised enrichment scores greater than 1.7 standard deviation were considered deregulated. Coherently, metabolites with uncorrected p-values smaller than 0.05 were considered deregulated. These values were chosen as they allow to generate a set of input of comfortable size to run CARNIVAL. Then, we first set the deregulated kinases, phosphatases and TFs as starting points and deregulated metabolites as end points (forward run). This direction represents regulations first going through the signaling and transcriptional part of the cellular network and stops at deregulated metabolites in the metabolic reaction network. However, since metabolite concentration can also influence the activity of kinases and TFs through allosteric regulations, we also ran CARNIVAL by setting deregulated metabolites as starting points and deregulated TFs, kinases and phosphatases as end points (backward run). For the forward run, after 7200 second of run time, CARNIVAL yielded a network of 76 edges, a feasible solution which proved to be within the 11.24% gap from the optimal. For the backward run, CARNIVAL found a solution within the 2.44% gap from the optimal after 7200 second of run time, yielding a network of 177 edges.

Since there were no incoherences in the predicted activity signs between the common part of the two resulting networks, they were simply merged together, resulting in a combined network of 250 unique edges.

4.6 Coherence between CARNIVAL mechanistic hypotheses and omics measurements

To assess the robustness of CARNIVAL predictions, we used two different methods. First, the CARNIVAL network contains cases where a protein activity is modelled by CARNIVAL as up- or down-regulated under the control of a TF. If such hypotheses are correct, then one would expect to see the abundance of the corresponding transcript of the proteins to be coherently up or down-regulated (since the control of the TF is carried through regulation of transcript abundance) (Supplementary Figure 3). Thus, a True Positive (TP) is defined as a carnival node that is directly downstream of a TF and has the same sign (-1 or 1) as a significantly deregulated corresponding transcript. Transcripts with LIMMA moderated absolute t-values ranging between 0 and 2 were considered as significantly deregulated. Since CARNIVAL predictions are discrete (-1, 0, 1), we can't make a classic receiving operator curve. Furthermore, we are lacking knowledge of True Negative. Indeed, node activities set to 0 by CARNIVAL cannot be interpreted because measurements and activities

inputs only cover a fraction of the PKN and consequently most of the PKN nodes will be set to 0 by default. We showed that the TPR were relatively stable between 0 and 1.7, and more volatile between 1.7 and 2, likely due to the number of significantly deregulated transcripts considered becoming too small (see Supplementary Figure 4A). To estimate the baseline TPR of a random algorithm, we consider the following question : If i take any transcript that was measured and randomly assign it a value of 1 (or -1), what is the probability that the transcript will indeed be significantly up-regulated (or down-regulated), for given a t-value threshold. This probability can be simply estimated from the actual proportion of transcripts that are significantly up-related. For absolute t-values ranging between 0 and 1.7 (number of transcripts = 13), carnival TPR was consistently higher than the random baseline, but performed equal or worse than random above 1.7, again likely due to the number of significantly deregulated transcripts considered becoming too small.

Second, when multiple nodes are co-regulated by a common parent node in the CARNIVAL network, we can assume that the activity of the co-regulated nodes should be correlated. Thus, we create a correlation network with the TF and kinase/phosphatase activities estimated at a single sample level. To estimate the single sample level activities, normalised RNA counts and phosphosite intensities were scaled (minus mean over standard deviation) across samples. Thus, the value of each gene and phosphosite is now a z-score relative to an empirical distribution generated from the measurements across all samples. We used these z-scores as input for the viper algorithm to estimate kinase/phosphatases and TF activities at single sample level. Thus, the resulting activity scores in a sample are relative to all the other samples. Then, a correlation network was built using only tumor samples. Thus, the correlation calculated this way represents co-regulations that are supported by the available data in tumor (number of coregulations = 157. We defined the ground truth for co-regulations as over a range of absolute correlation coefficients between 0 and 1 with a 0.01 step. Thus, a True Positive here is a co-regulation predicted from the topology of the carnival network that also has a corresponding absolute correlation coefficient in tumor samples above the given threshold. Since defining a ground truth in such a manner can yield many false positives (a correlation can often be spurious), the TPR of COSMOS was always compared to a random baseline.

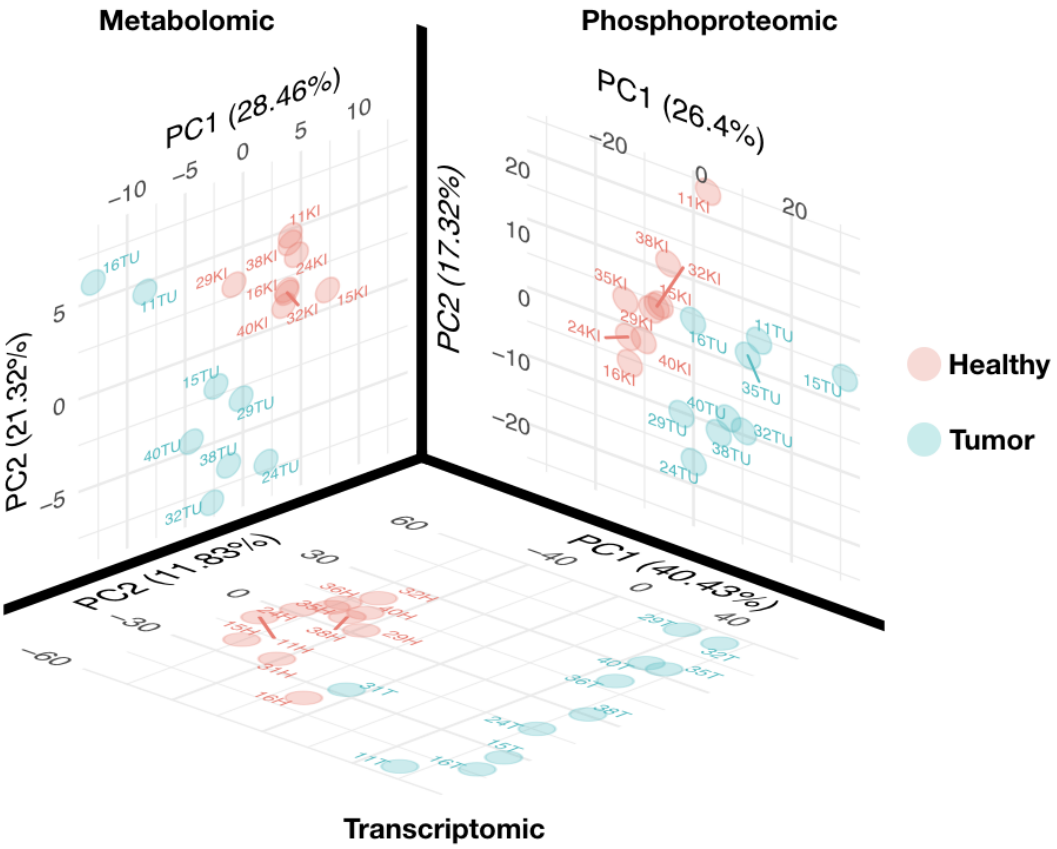
4.7 Code availability

All code used in this study is available at : https://github.com/saezlab/cosmos_prototype

4.8 Data availability

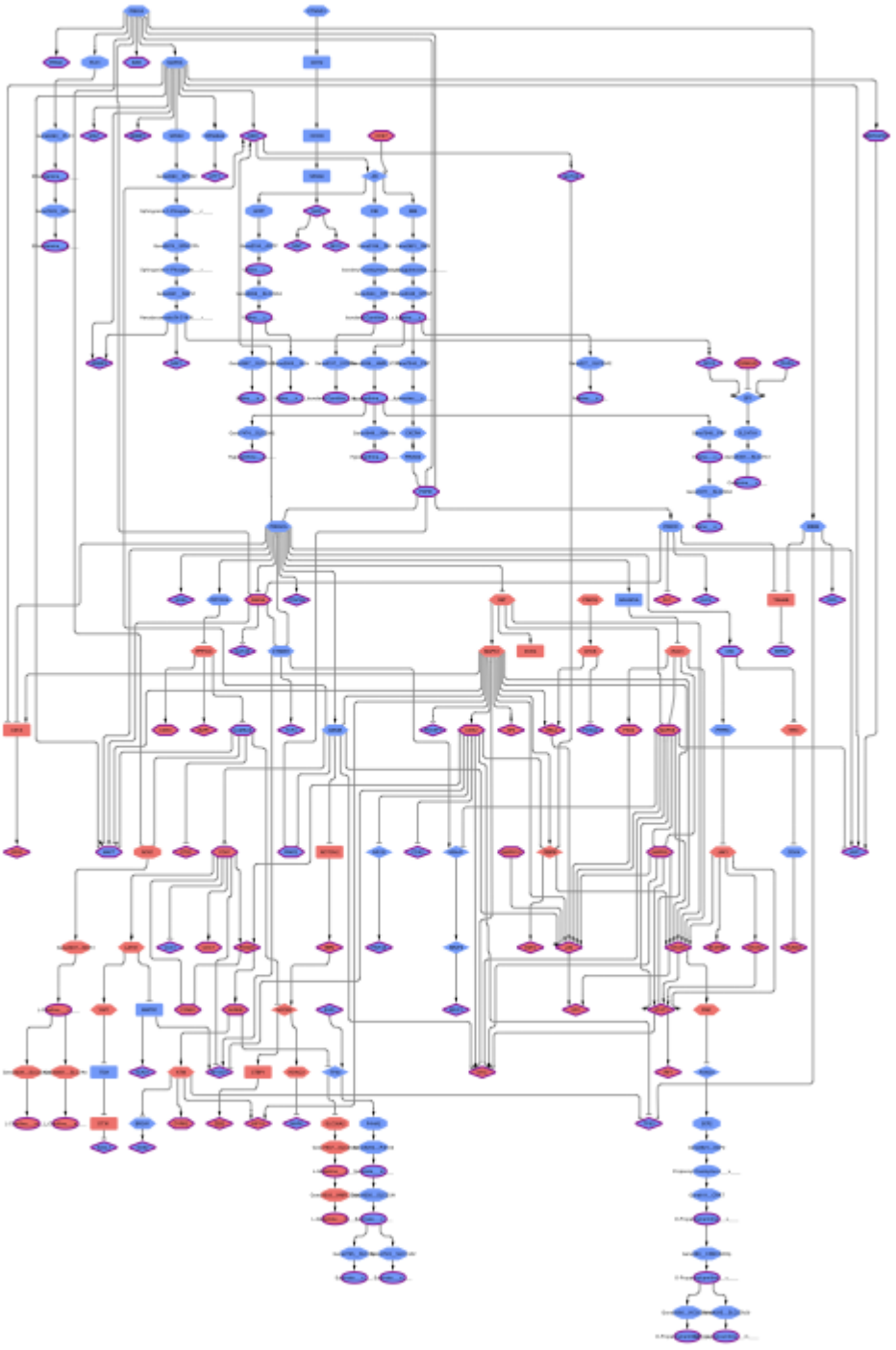
Data used in this study is available at : https://github.com/saezlab/cosmos_prototype/tree/main/data

Supplementary Materials



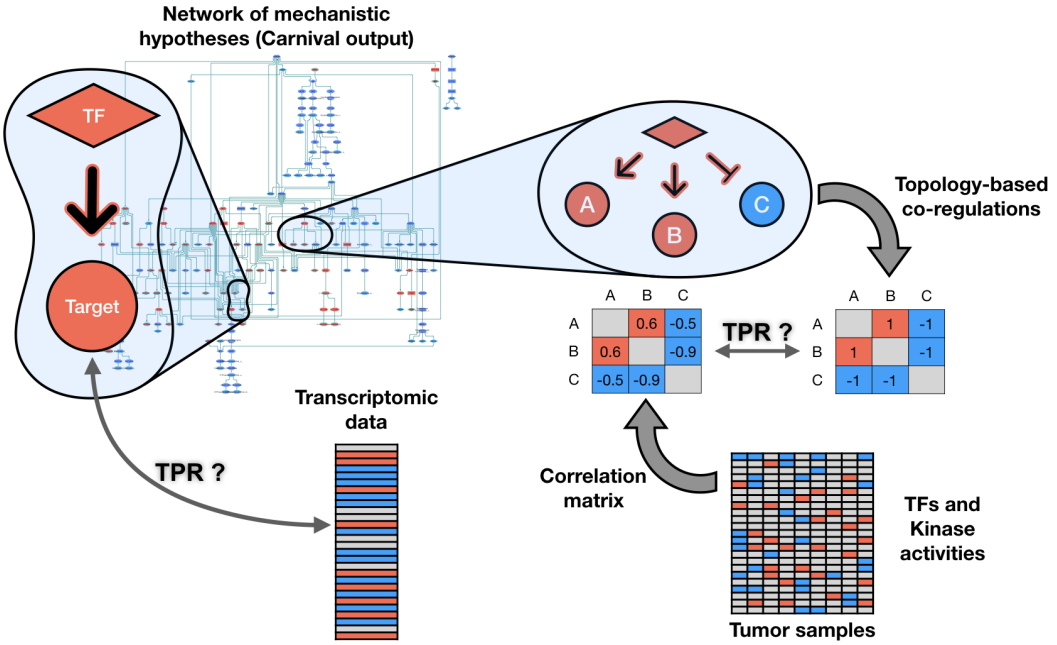
Supplementary Figure 1

PCA of Metabolomics, Phosphoproteomics and transcriptomics datasets for tumor and healthy tissues samples. For each omics dataset, PCA is run independently on normalised datasets and the first two components are plotted. Each omics shows a clear separation between tumor and healthy tissue.



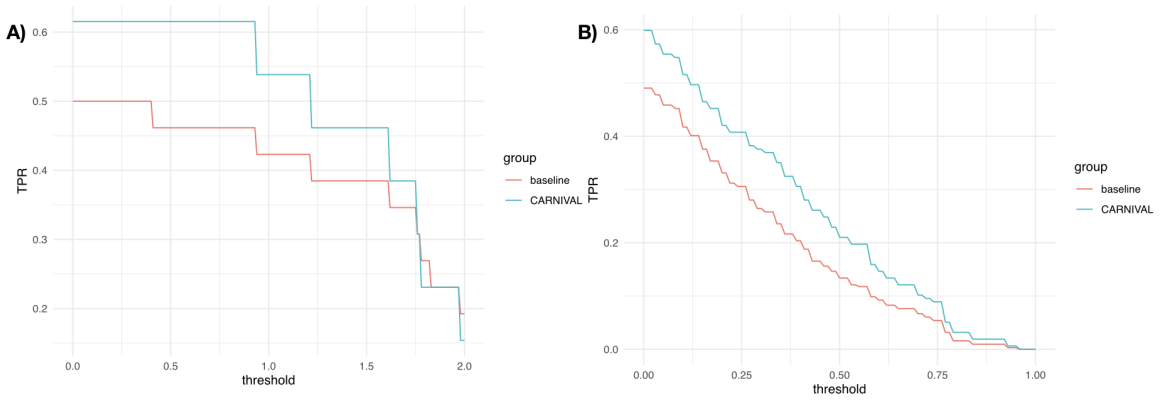
Supplementary Figure 2

Causal network summarising the mechanistic hypotheses systematically generated by CARNIVAL. (see Figure 4 for legend). The network comprises 250 edges. It represents the propagation of signals connecting the deregulated kinases, phosphatases, TFs and metabolites in kidney cancer.



Supplementary Figure 3

Coherence assessment between CARNIVAL hypotheses and underlying data. On the left, the predicted activity TF targets of the COSMOS network are compared to the actual t-value (tumor - healthy) of their corresponding transcript to determine true positive rate (TPR). On the right, coregulations predicted by COSMOS are compared against a correlation network of kinase/TF activities to determine TPR.



Supplementary Figure 4

Exploration of TPR stability in function of the chosen t-value/correlation threshold. A) For TF/transcriptomics data coherence. True positive rates are estimated over a range of t-value between 0 and 2 with 0.1 steps. B) For the correlation/topology coherence. True positive

rates are estimated over a range of Pearson correlation between 0 and 1 with 0.01 steps. In A) and B) COSMOS (CARNIVAL) performance is compared to a random baseline. COSMOS consistently outperforms the random baseline.

Chapter 4 : Metabolic enzyme footprint analysis

Abstract

The functional insights that metabolomic data sets contain currently lies under-exploited. This is in part due to the complexity of metabolic reaction networks and the indirect relationship between reaction fluxes and metabolite abundance. Yet, footprint-based methods have been available for decades in the context of other omic data sets such as transcriptomic and phosphoproteomic. Here, we present ocEAn, a method that defines metabolic enzyme footprint from a curated reduced version of the recon2 reaction network and use them to explore coordinated deregulations of metabolite abundances with respect to their position relative to metabolic enzymes in the same manner as Kinase-substrate and TF-targets Enrichment analysis. We show how ocEAN can consistently help identify deregulated metabolic enzyme activities by comparing its output with proteomic data and a chemical enzyme inhibition experiment.

1. Introduction

The signaling machinery of cells has evolved with the prime goal of allowing ancestral single cell organisms to adapt their energy metabolism to shifting environmental conditions. This effectively places the study of metabolism at the heart of understanding cell biology. Ultimately, even an organism as complex as a human being can be conceptualised as a population of cells differentiating and cooperating to secure an overall constant intake of energetic substrates (I eat, therefore I am, therefore I eat, etc..). Coherently, many diseases and chronic health conditions, such as cancer and kidney fibrosis, are associated with striking cellular metabolic reprogramming ((Chen & Xiong, 2020; Cocetta *et al*, 2020; Sulkowski *et al*, 2020)). A common way of studying cell metabolism is to measure the abundance of dozens to hundreds of metabolites at the same time in a tissue/cell culture using mass-spectrometry coupled with liquid chromatography (LC-MS) or gas chromatography(GC-MS), generating what is referred to as “metabolomic” data sets. Despite an ever growing body of available metabolomic datasets, extracting functional insights out of those remains very challenging, likely due partially to the very complex and nonlinear nature of their underlying metabolic reaction networks.

In this context, we sought to translate the footprint based analysis we had worked extensively on with transcriptomic and phosphoproteomic data(Garcia-Alonso *et al*, 2019; Holland *et al*, 2020; Hernandez-Armenta *et al*, 2017; Terfve *et al*, 2015) to metabolomic data. The goal of such an approach is to estimate the activity change of a metabolic enzyme by integrating the metabolic abundances changes happening downstream and upstream of its position in a complex metabolic reaction network. It is widely accepted that blocking/increasing the activity of a specific metabolic enzyme results in accumulation/depletion of its reactants and/or a symmetrical depletion/accumulation of its products, respectively. This concept was the ground assumption of the reporter reaction(Cakir *et al*, 2006). The reporter reaction method was trying to identify reactions that were in the middle of coordinated (up/down) metabolic abundance deregulations. This hypothesis was later further supported by additional experimental evidence(Ewald *et al*, 2013). These metabolic abundances changes are also expected to propagate to a certain extent to other up and downstream metabolites. Thus, a more recent method called metabolic network segmentation (MNS) aimed to exploit this assumption to find key deregulated metabolic enzymes in the reaction network by integrating topological clusters of coordinated metabolic abundance deregulations(Kuehne *et al*, 2017). It essentially used the topology of the network reaction to define groups of metabolites based on their proximity in the network and how well coordinated their abundance changes were. However, MNS only highlights pivotal (right in between clusters of up and down-regulated metabolites) reactions.

In this study we present ocEAn (Metabolic enzyme Metabolite Set enrichment analysis), a method based on footprint-based activity estimation such as TFEA and KSEA (see 4. Footprint analysis) and exploiting the same metabolic assumption as the Reporter Reaction and NMS methods. It has the advantage over NMS to report scores of activity for all enzymes of a metabolic reaction network. It also relies on a curated human metabolic network, which allows more accurate estimations. Finally, it also provides activity score estimations that are coherent with the ones provided by other footprint-based methods. We applied ocEAn on a metabolomics dataset generated from a kidney cancer cell line model (786-O) and validated our findings using proteomic data, metabolic labelling experiments and targeted inhibition of metabolic enzymes.

2. Results and methods

2.1 Causal format of reduced recon2 human metabolic reaction model

The first thing we need to run ocEAn on any type of metabolomic data are the sets of metabolites that are associated with each metabolic enzyme. This information can be extracted from the metabolic reaction network, which informs us on which metabolite is downstream or upstream of each metabolic reaction. Thus, we used a reduced manually curated and thermodynamically proofed version of the Recon2 human metabolic reaction network to generate the metabolite sets. The thermodynamic proofing was performed to exclude reaction directions that were not thermodynamically feasible with the TFBA algorithm (Kiparissides & Hatzimanikatis, 2017).

In order to get a relevant idea of the relative position of metabolites with respect to enzymes, it is also important to filter out accessory elements of the reaction network such as cofactors and over-promiscuous metabolites. Over-promiscuous metabolites are metabolites that are used as reactants by a very large number of reactions. Thus, changes in over-promiscuous metabolites abundances hold little discriminating power to estimate metabolic enzyme activities. Furthermore, they are often not the main reactant of a reaction and will create many irrelevant bridges between unrelated reactions in the network. For that, metabolites classified as cofactors and nucleotides according to the KEGG BRITE classification were removed, as well as CO2, ITP, IDP, NADH and all metabolites composed of less than 4 atoms. This procedure effectively filtered out 100 metabolites, bringing the number of metabolites in the reaction network from 421 to 321. The network was then “causalised” using the same procedure as described in 4.4 Meta PKN construction. The resulting causal reaction network allows to easily follow paths connecting metabolic enzymes with distant metabolites. The next step consist in associating each enzyme of the network and all metabolites of the network with weights representing the minimum distance of metabolites relative to enzymes and a sign representing whether a each metabolite is upstream (-1) or downstream (1) of a given enzyme (See Figure 1).

In order to compute a weight, we used a function that progressively decreases a weight value. The weight value starts at 1 for direct reactant and products of a given enzyme and decrease in a stepwise manner ($x_{i+1} = x_i$ * penalty, with $x_0 = 1$ and penalty ranging between 0 and 1), for each reaction step separating the given metabolite from a given enzyme. The range is defined between 0 and 1 to yield a ‘contribution of metabolite statistic’ to the final

score ranging between 0 to 100% of their respective abundance change. The penalty (ranging between 1 and 0) will determine how fast a metabolite loses its influence on the score of a given enzyme with increasing distance in the metabolic reaction network (Figure 2).A penalty set to 0 would mean that only direct reactants and products are taken into account when estimating an enzyme activity score (which would correspond to the Reporter Reaction method). A penalty of 1 would mean that all metabolites that are exclusively upstream or downstream of an enzyme are taken into account equally.

Since a lot of cycles are present in the metabolic reaction network, metabolites are usually both upstream and downstream of enzymes. To recover a weight that represents the actual relative position of a metabolite with respect to a given enzyme, the upstream and downstream weight of each metabolite-enzyme associations are averaged.

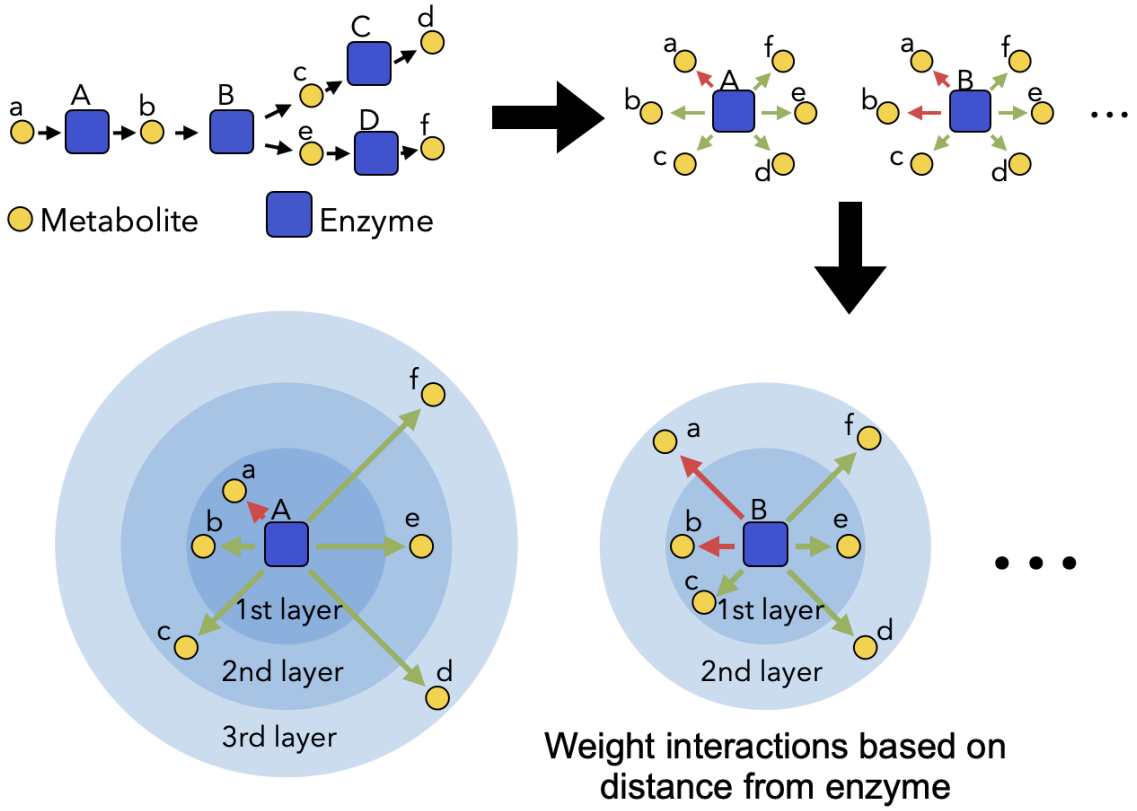


Figure 1

Schematic of the procedure applied to convert a causal reaction network into sets of metabolites.

The resulting metabolite sets are then ready to be used with enrichment algorithms such as viper to estimate metabolic enzyme activities from metabolomic data sets, as is done with

phosphoproteomic for kinase activities and transcriptomic for transcription factor activities (4. Footprint analysis).

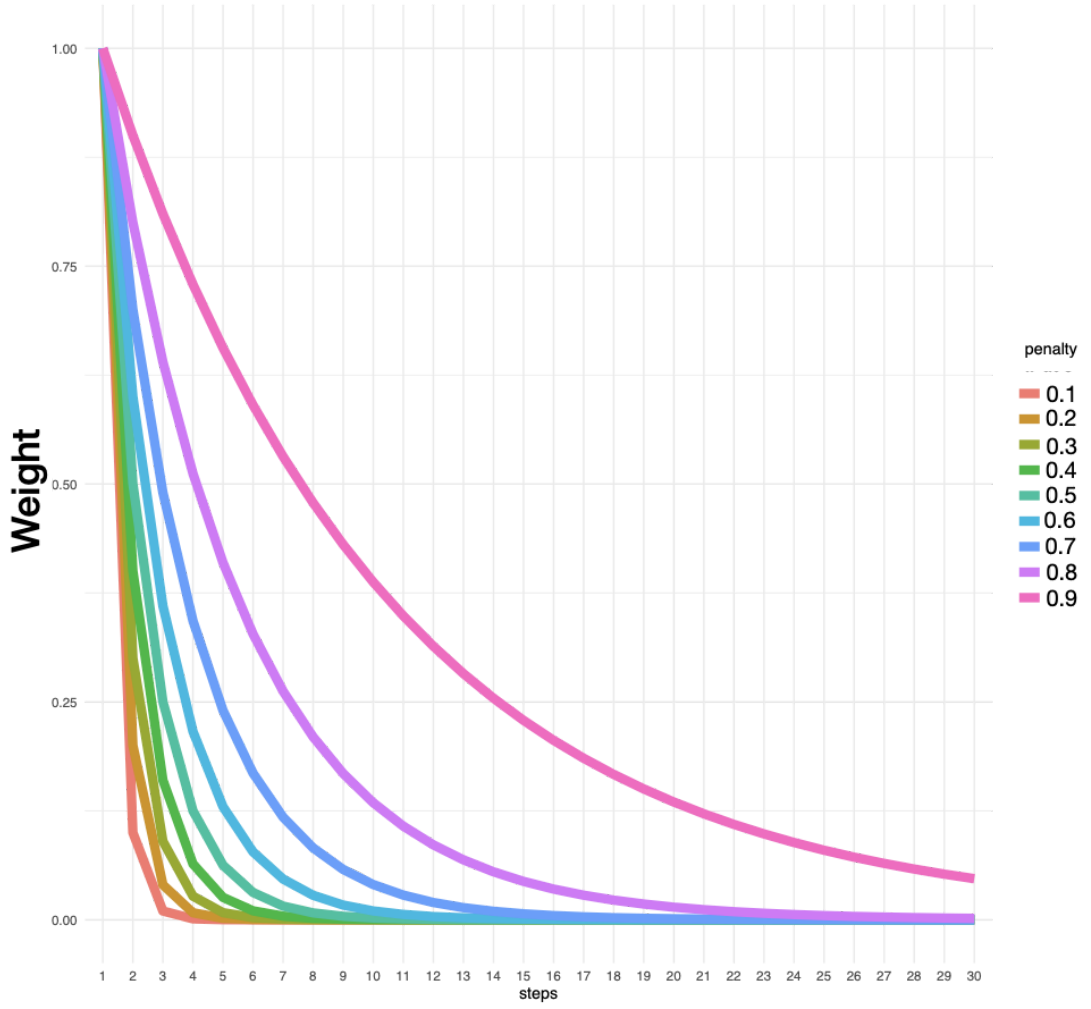


Figure 2

Decrease of weight with increasing distance (reaction steps) between a metabolite and a given enzyme for different penalty values.

2.2 Systematic metabolic enzyme activity estimation

In order to test the ocEAn method, we applied it on a metabolomic data set generated from HK2 and 786-O cell lines. HK2 cells are immortalised kidney cells while 786-O are cells derived from a kidney tumor. Thus, in order to study the potential metabolic reprogramming occuring in kidney cancer, we performed a limma differential analysis of 76 cellular metabolites detected across 17 biological replicates of HK2 and 17 biological replicates of 786-O. The t-values resulting from the limma differential analysis were used alongside the metabolite sets generated above with the viper algorithm to estimate metabolic enzyme activity changes in 786-O compared to HK2. The distance penalty of the metabolite set was

set to 0.6, in order to give a moderate importance to metabolites situated far away from enzymes. Indeed, it is hard to know which value for this parameter will really yield the best results, and it is very likely context dependent (coverage of the metabolic reaction network, which part of the metabolic reaction network is best covered, etc...). Thus, a cutoff of 0.6 seemed like a sensible intermediate value. This choice and the impact of the full range of penalties between 0.1 and 1 in this specific analysis context are studied in the next section (2.3).

This analysis yielded activity estimation for 726 unique enzyme and metabolic enzyme complexes in 3580 unique reactions (each enzyme is associated with multiple different reactions and directions). The results were further interpreted biologically and yielded interesting insights such as a drainage of mitochondrial metabolites (through BCAT2, Supplementary figure 1). It also hinted at a reprogramming of branching amino-acid metabolism to provide Alpha-keto-glutarate (Akg) and aspartate to support the heavy nucleotide synthesis activity of cancer cells through rewiring of branching amino-acid metabolism (through BCAT1, GOT and MDH family enzymes, Supplementary figure 1). These results are being further explored in the context of a manuscript currently in preparation, dissecting the metabolic landscape of kidney tumor progression.

2.3 Comparison of metabolic enzyme activity with proteomic data and validation

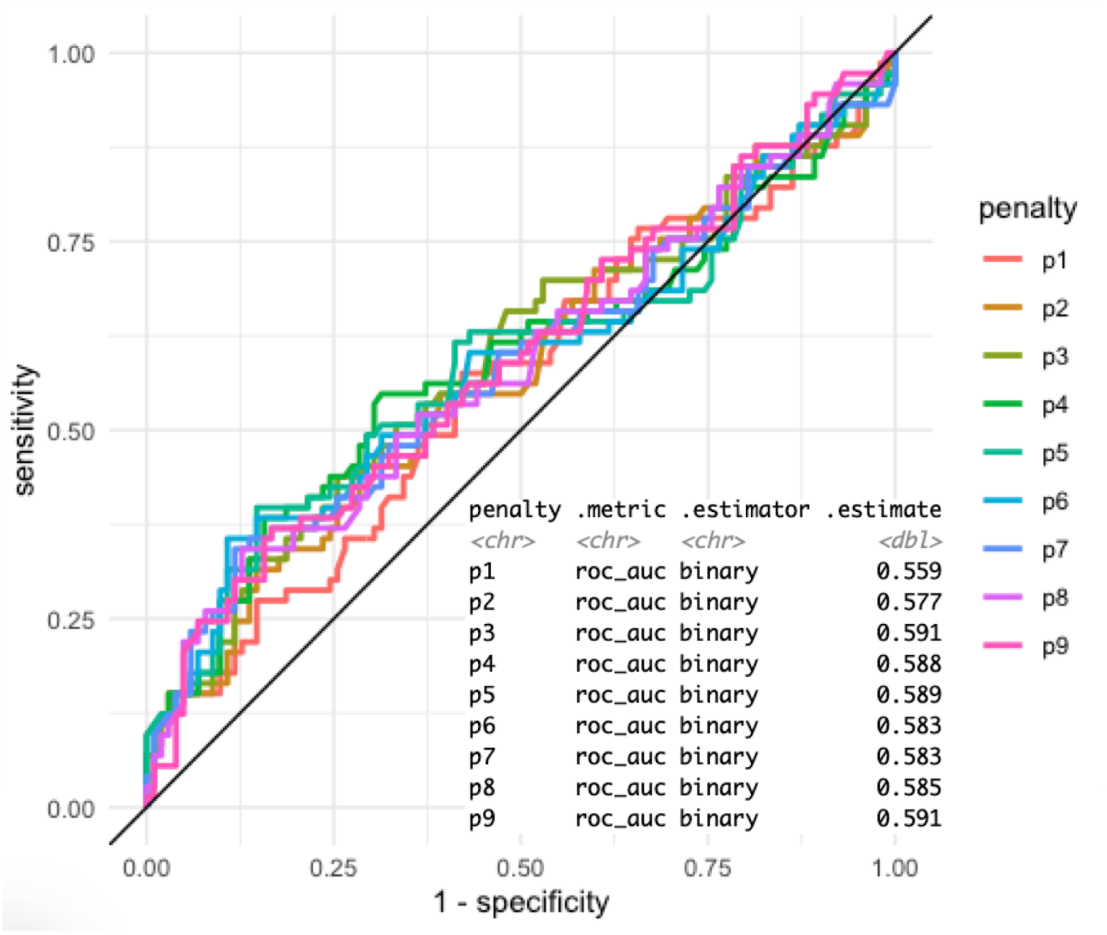
It is quite challenging to estimate the actual performance of a metabolic enzyme activity estimation tool as the ground truth corresponding to its prediction is very hard to access. The activity output of a metabolic enzyme is essentially how many reactants are converted in products in a given interval. This is often referred to as a metabolic enzyme flux. The net flux of a metabolic enzyme is the sum of it's fluxes in both directions and is what is usually measured. Experimentally measured net fluxes at the scale of a few selected reactions is already difficult to estimate, and currently impossible to obtain in a systematic manner. In order to get a rough idea of the performances of ocEAn, we compared its result to enzyme abundances changes obtained from proteomic data and to a chemical inhibition of the BCAT1 metabolic enzyme data set. We chose BCAT1 because it was consistently predicted to be up-regulated in 786-O compared to HK2 by ocEAn.

First, limma differential analysis was performed on a proteomic dataset generated from the same cell lines as previously (786-O and HK2). Then a Receiving Operator Curve (ROC) and Precision Recall Curve (PRC) analysis was performed to systematically compare the

ocEAn output activity estimations with 1) the significant positive proteomic abundance changes (t-value > 1.7) and 2) the significant negative proteomic abundance changes (t-value < -1.7). For 1) positive changes, his showed that the best area under the ROC (AUROC) was obtained when the penalty was set to 0.3 or 0.9, which corresponds to the maximum weights given to far away metabolites out of all tested penalties. However, the AUROC values were not very different from penalty 0.3 to 0.9 (0.58 to 0.59 in positive changes, 0.58 to 0.61 for negative changes) (Figure 3, Figure 4). This also hinted that ocEAn's output seems slightly more consistent with negative abundance changes than positive ones. The PRC analysis showed similar results, with 1) positive changes comparison yielding best area under PRC (AUPRC) value for the penalty value 0.5, and 2) negative changes comparison yielding best AUPRC values for penalties between 0.6 and 0.9. These AUROC are consistent with the expectation that only a minor part of the variance of an enzyme activity is actually explained by its abundance. Indeed, we expect to see many cases where a change of metabolic enzyme activity will be inconsistent with the direction of its protein abundance change. Thus, an AUROC value of 1 is not the expected goal in this analysis. Furthermore, this analysis was only performed in the context of one comparison between two conditions. These results made it apparent that a clear answer to which is the best penalty value may need further studies. Thus, a distance penalty of 0.6 for further analysis seemed a reasonable choice at this point, as it falls between the best AUROC and AUPRC values. The same analysis was performed using transcriptomic data instead of proteomic. It showed that ocEAn outputs were much less consistent with transcriptomic changes than with proteomic ones. Indeed, for both positive and negative abundance changes, the AUROC values remained around 0.5 for all penalties, while the AUPRC values were barely above the random baseline (Supplementary figure 2). All code to reproduce the AUROC and AUPRC analysis can be found at: https://github.com/saezlab/ocean_thesis

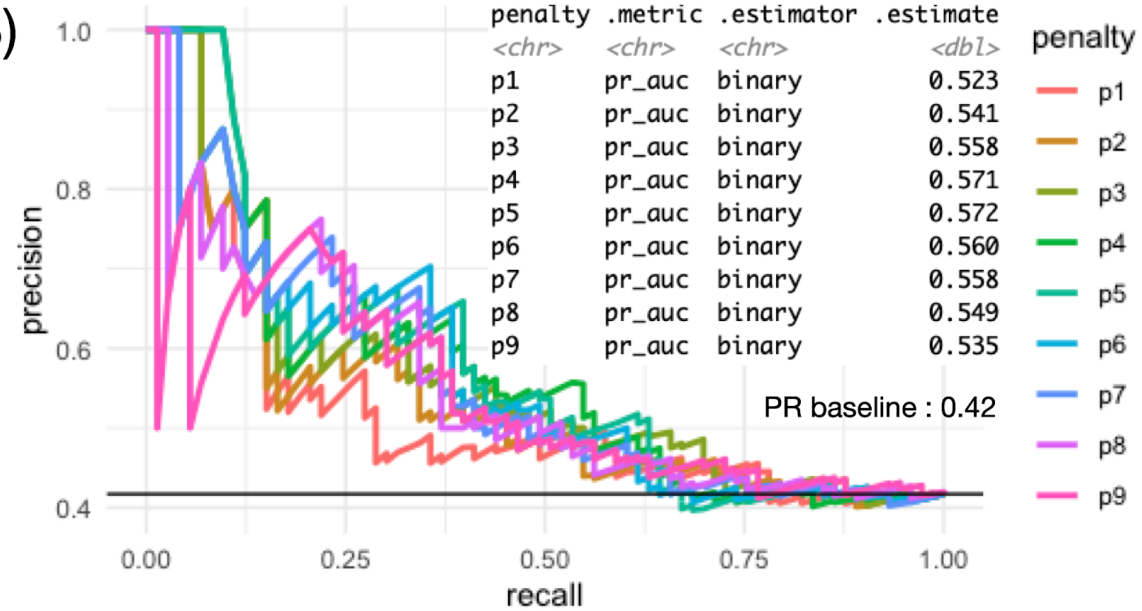
Then, another metabolomic data set was generated from 786-O and HK2 cell lines were the BCAT1 and 2 enzymes were chemically inhibited. This enzyme was chosen as it is a key enzyme of branching amino acid metabolism, a pathway that is known to be severely deregulated in kidney cancer. Furthermore, the BCAT2 isoforms of the enzymes was predicted to be strongly up-regulated by ocEAn in 786-O compared to HK2. ocEAn predicted consistently a downregulation of the activity of BCAT2 in the BCAT inhibited 786-O cells as well as in HK2 but to a much milder degree. This is consistent with the expected result since BCAT1 and 2 are thought to be much more active in the 786-O cell than in HK2.

A)



penalty	.metric	.estimator	.estimate
<chr>	<chr>	<chr>	<dbl>
p1	roc_auc	binary	0.559
p2	roc_auc	binary	0.577
p3	roc_auc	binary	0.591
p4	roc_auc	binary	0.588
p5	roc_auc	binary	0.589
p6	roc_auc	binary	0.583
p7	roc_auc	binary	0.583
p8	roc_auc	binary	0.585
p9	roc_auc	binary	0.591

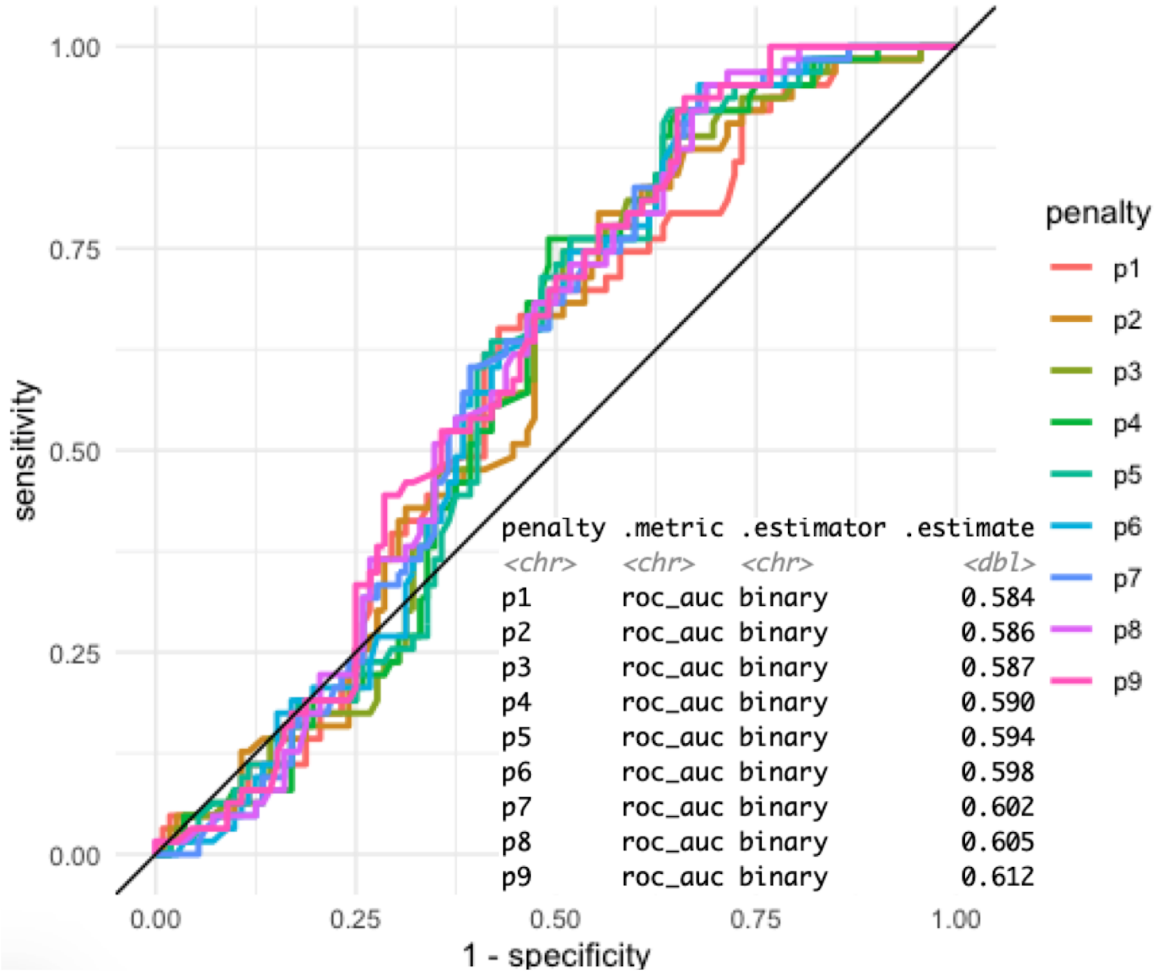
B)



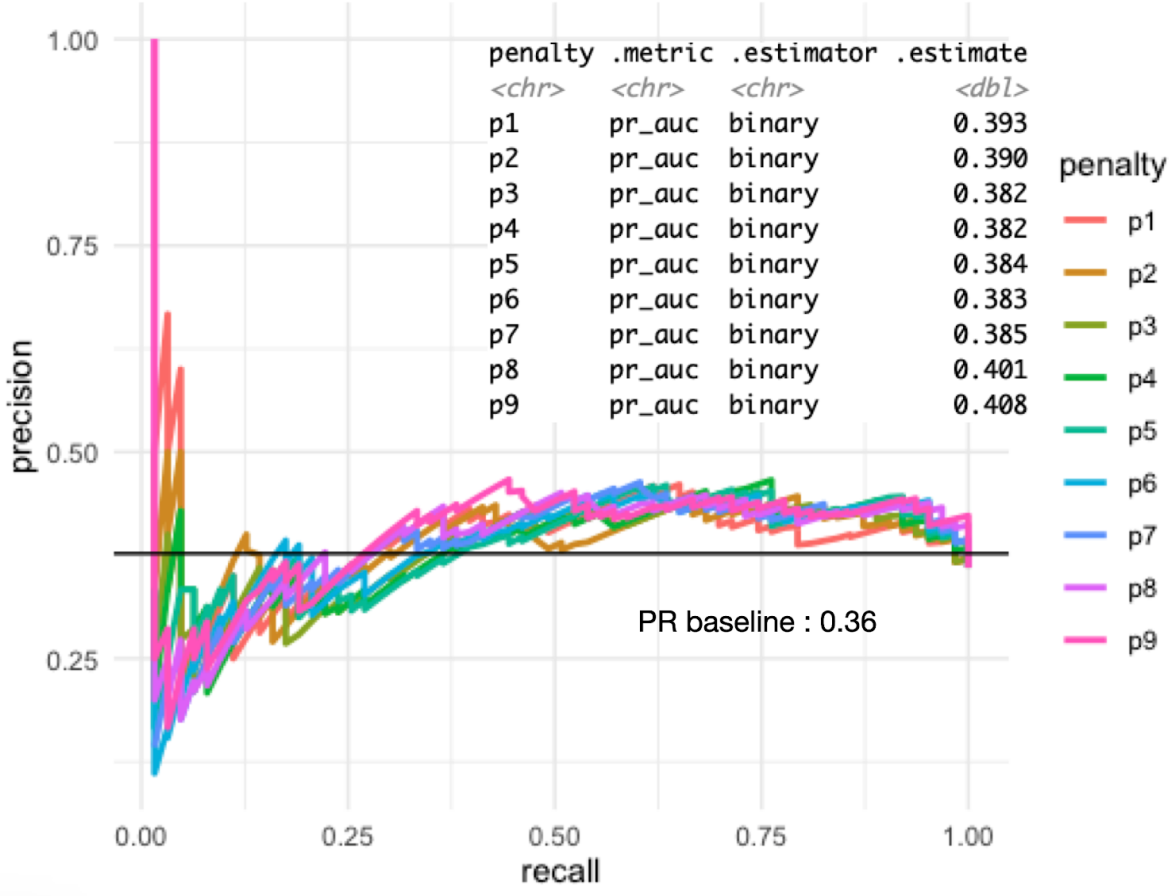
penalty	.metric	.estimator	.estimate
<chr>	<chr>	<chr>	<dbl>
p1	pr_auc	binary	0.523
p2	pr_auc	binary	0.541
p3	pr_auc	binary	0.558
p4	pr_auc	binary	0.571
p5	pr_auc	binary	0.572
p6	pr_auc	binary	0.560
p7	pr_auc	binary	0.558
p8	pr_auc	binary	0.549
p9	pr_auc	binary	0.535

Figure 3

A) AUROC and B) AUPRC for the comparison between up-regulated protein abundances with ocEAn metabolic enzyme activity estimations for a range of penalty values between 0.1 (p1) and 0.9 (p9).



penalty	.metric	.estimator	.estimate
<chr>	<chr>	<chr>	<dbl>
p1	roc_auc	binary	0.584
p2	roc_auc	binary	0.586
p3	roc_auc	binary	0.587
p4	roc_auc	binary	0.590
p5	roc_auc	binary	0.594
p6	roc_auc	binary	0.598
p7	roc_auc	binary	0.602
p8	roc_auc	binary	0.605
p9	roc_auc	binary	0.612



penalty	.metric	.estimator	.estimate
<chr>	<chr>	<chr>	<dbl>
p1	pr_auc	binary	0.393
p2	pr_auc	binary	0.390
p3	pr_auc	binary	0.382
p4	pr_auc	binary	0.382
p5	pr_auc	binary	0.384
p6	pr_auc	binary	0.383
p7	pr_auc	binary	0.385
p8	pr_auc	binary	0.401
p9	pr_auc	binary	0.408

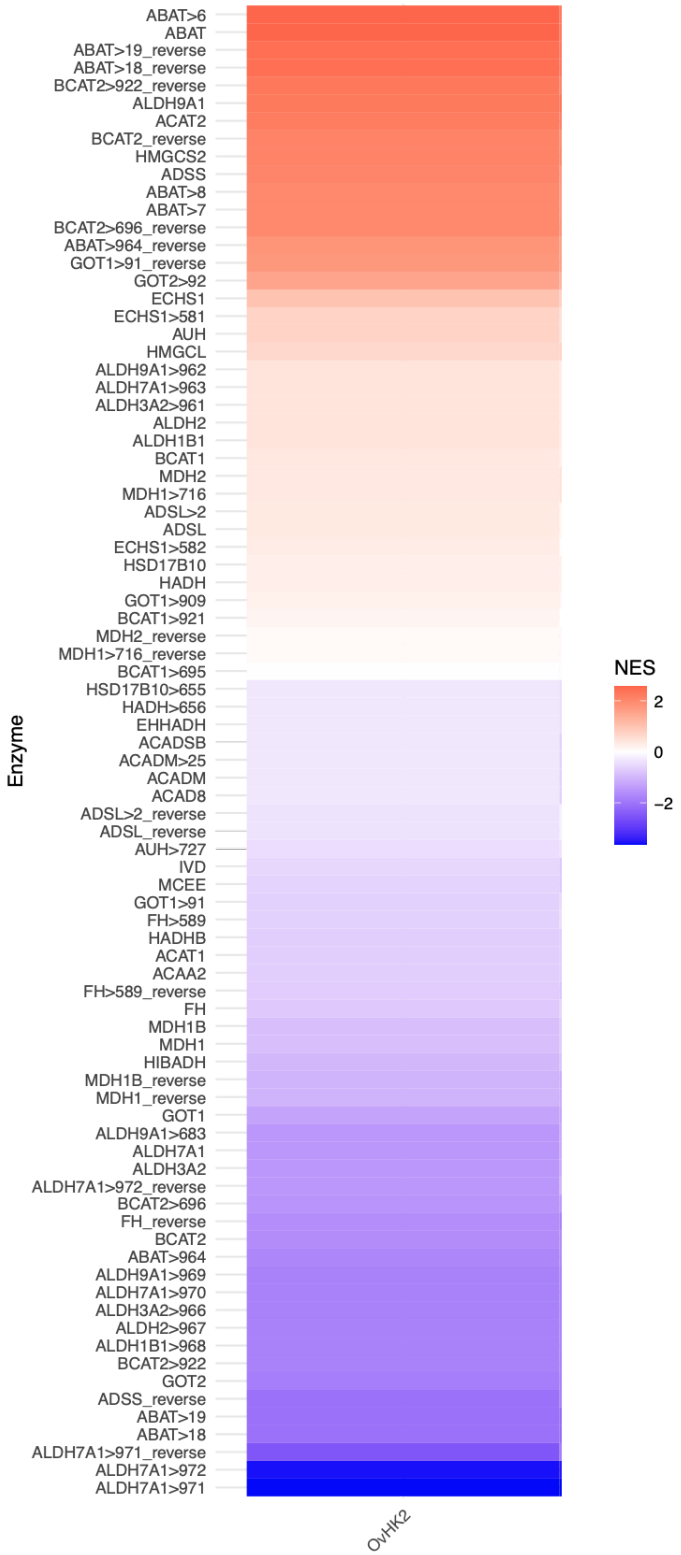
Figure 4

A) AUROC and B) AUPRC for the comparison between down-regulated protein abundances with ocEAn metabolic enzyme activity estimations for a range of penalty values between 0.1 (p1) and 0.9 (p9).

3. Discussion

In this chapter, we showed how footprint based methods designed for transcriptomic and phosphoproteomic can be translated to metabolomic data by adapting the specificities of metabolic reaction networks to generate metabolic enzyme footprints. The metabolic enzyme footprints were then used as metabolite sets to perform metabolic enzyme enrichment analysis in a similar fashion as Kinase-substrate/TF enrichment analysis with a method called ocEAn. ocEAn seems to be performing relatively well to estimate metabolic enzyme activity when compared to proteomic data or chemical inhibition of specific metabolic enzymes. It can be theoretically applied to any metabolomic dataset generated from human cells. It is currently being developed further in the context of an analysis of the metabolic landscape of kidney cancer. In the future, it will be interesting to see how much of the quality of ocEAn estimates are dependent on the quality of the metabolic reaction network used to generate the set of metabolite-enzyme distances. Indeed, I suspect that an adequate prior-knowledge source is usually what impacts the quality of footprint-based activity estimations the most, which will need to be further investigated. Finally, since ocEAn yields metabolic enzyme activities that are conceptually similar to TF and Kinase activities estimated with TFEA and KSEA, a logic future step could consist in integrating the ocEAn metabolic enzyme activities with TF and kinase activities across multiple omic layers. This could be done for example by connecting metabolic enzyme, TF and kinase activities with tools such as COSMOS.

4. Supplementary figures



Supplementary figure 1

Activity estimations for branching amino acid metabolism related metabolic enzymes. The activity score is a normalised enrichment score estimated with a weighted mean normalised through metabolite shuffling.

A) Comparison between up-regulated transcripts and ocEAn metabolic enzyme activity estimations

penalty	.metric	.estimator	.estimate
<chr>	<chr>	<chr>	<dbl>
p1	roc_auc	binary	0.535
p2	roc_auc	binary	0.522
p3	roc_auc	binary	0.514
p4	roc_auc	binary	0.509
p5	roc_auc	binary	0.505
p6	roc_auc	binary	0.501
p7	roc_auc	binary	0.498
p8	roc_auc	binary	0.498
p9	roc_auc	binary	0.495

penalty	.metric	.estimator	.estimate
<chr>	<chr>	<chr>	<dbl>
p1	pr_auc	binary	0.366
p2	pr_auc	binary	0.386
p3	pr_auc	binary	0.372
p4	pr_auc	binary	0.406
p5	pr_auc	binary	0.385
p6	pr_auc	binary	0.418
p7	pr_auc	binary	0.419
p8	pr_auc	binary	0.394
p9	pr_auc	binary	0.408

PR baseline : 0.4

B) Comparison between down-regulated transcripts and ocEAn metabolic enzyme activity estimations

penalty	.metric	.estimator	.estimate
<chr>	<chr>	<chr>	<dbl>
p1	roc_auc	binary	0.525
p2	roc_auc	binary	0.519
p3	roc_auc	binary	0.513
p4	roc_auc	binary	0.514
p5	roc_auc	binary	0.516
p6	roc_auc	binary	0.517
p7	roc_auc	binary	0.514
p8	roc_auc	binary	0.511
p9	roc_auc	binary	0.508

penalty	.metric	.estimator	.estimate
<chr>	<chr>	<chr>	<dbl>
p1	pr_auc	binary	0.519
p2	pr_auc	binary	0.529
p3	pr_auc	binary	0.520
p4	pr_auc	binary	0.521
p5	pr_auc	binary	0.515
p6	pr_auc	binary	0.505
p7	pr_auc	binary	0.494
p8	pr_auc	binary	0.483
p9	pr_auc	binary	0.472

PR baseline : 0.5

Supplementary figure 2

A) AUROC and AUPRC for the comparison between up-regulated transcript abundances with ocEAn metabolic enzyme activity estimations for a range of penalty values between 0.1 (p1) and 0.9 (p9). D) AUROC and AUPRC for the comparison between down-regulated transcript abundances with ocEAn metabolic enzyme activity estimations for a range of penalty values between 0.1 (p1) and 0.9 (p9).

Final conclusion

Cancer and fibrosis can be referred to as systemic diseases, due to the broad range of effects they can have on their host. With the ever increasing amount of prior-knowledge generated by the scientific community and along with the sharpening of measurement tools resolution, systematic analysis methods are a very powerful approach to extract relevant information from omics data sets. Indeed, they help to find which are the relevant mechanisms that govern the deep phenotypic reprogramming that can occur in the diseased cells. In this thesis, I have presented how footprint-based analysis and prior knowledge guided causal reasoning can be used to analyse large 'omics' biological data sets. Footprint-based methods have proven particularly useful to extract functional insights from measurements of abundances. I have especially shown that these functional insights were crucial in order to be able to connect multiple omic layers together with causal networks. Indeed, while different types of omics will focus on different types of molecules, such as RNAs, proteins and metabolites, footprint-based enzyme activities allow to bring back this data to more homogeneous features, e.g. transcription factors and kinases. With COSMOS, I presented the first attempt to systematically connect such TF, kinases and metabolic together with a global causal reasoning approach. This type of approach is helpful to connect together cellular processes spanning across multiple compartments and functions. For example, connecting signaling and metabolism together with such an approach can prove particularly useful to understand deregulation happening in complex multifactorial diseases such as cancer and fibrosis.

Parallel to this, metabolomic data still hold a great depth of under-exploited information. In this context, I developed ocEAn to systematically extract relevant functional information from metabolomic datasets. A key aspect of ocEAn metabolic enzyme activity estimations is that these functional outputs could then be connected to other omic layers, in the same way as it was done with TFs and kinases. For this reason, I hope that in the future these metabolic enzyme activities could be further integrated with tools like COSMOS.

Thus, COSMOS and ocEAn can help to pin-point relevant pathways and biological molecules that can serve as disease markers or therapeutic targets. They may also serve as a groundwork to provide functional insights from omic datasets that could then further be used to connect processes across multiple cell types and tissues. With the help of pan-tissue and pan-cell type prior knowledge networks, we can hold the hope of building descriptive

disease models recapitulating their broad effect over an entire body. We will also increase the resolution of our tools to a single cell resolution data set, in an effort to generate mechanistic insights from the whole organism to single cell scales. While there is still a long way before reaching this point, the direction technology is currently evolving makes this goal appear as more and more realistic every day.

"Imaginary mountains build themselves from our efforts to climb them, and it's our repeated attempt to reach the summit that turns those mountains into something real."

Bennett Foddy

References

Ackermann M & Strimmer K (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10: 47

Alhamdoosh M, Ng M, Wilson NJ, Sheridan JM, Huynh H, Wilson MJ & Ritchie ME (2017) Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics* 33: 414–424

Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH & Califano A (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* 48: 838–847

Amadoz A, Hidalgo MR, Çubuk C, Carbonell-Caballero J & Dopazo J (2018) A comparison of mechanistic signaling pathway activity analysis methods. *Brief Bioinform*

Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W & Stegle O (2018) Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 14: e8124

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29

Auslander N, Yizhak K, Weinstock A, Budhu A, Tang W, Wang XW, Ambs S & Ruppin E (2016) A joint analysis of transcriptomic and metabolomic data uncovers enhanced enzyme-metabolite coupling in breast cancer. *Sci Rep* 6: 29662

Ayati M, Wiredja D, Schlatzer D, Maxwell S, Li M, Koyuturk M & Chance M (2018) CoPhosK: A Method for Comprehensive Kinase Substrate Annotation Using Co-phosphorylation Analysis. *bioRxiv*: 251009

Berndt N, Egners A, Mastrobuoni G, Vvedenskaya O, Fragoulis A, Dugourd A, Bulik S, Pietzke M, Bielow C, van Gassel R, *et al* (2020) Kinetic modelling of quantitative proteome data predicts metabolic reprogramming of liver cancer. *Br J Cancer* 122: 233–244

Bertran-Alamillo J, Cattan V, Schoumacher M, Codony-Servat J, Giménez-Capitán A, Cantero F, Burbridge M, Rodríguez S, Teixidó C, Roman R, *et al* (2019) AURKB as a target in non-small cell lung cancer with acquired resistance to anti-EGFR therapy. *Nat Commun* 10: 1812

Bouhaddou M, Memon D, Meyer B, White KM, Rezelj VV, Correa Marrero M, Polacco BJ, Melnyk JE, Ulferts S, Kaake RM, *et al* (2020) The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell* 182: 685–712.e19

Bradley G & Barrett SJ (2017) CausalR: extracting mechanistic sense from genome scale data. *Bioinformatics* 33: 3670–3672

Brenner W, Färber G, Herget T, Wiesner C, Hengstler JG & Thüroff JW (2003) Protein kinase C eta is associated with progression of renal cell carcinoma (RCC). *Anticancer Res* 23: 4001–4006

Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, Gatto F, Nilsson A, Preciat Gonzalez GA, Aurich MK, *et al* (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat Biotechnol* 36: 272–281

Cakir T, Patil KR, Onsan ZI, Ulgen KO, Kirdar B & Nielsen J (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. *Mol Syst Biol* 2: 50

Cantini L, Calzone L, Martignetti L, Rydenfelt M, Blüthgen N, Barillot E & Zinovyev A (2018)

Classification of gene signatures for their information value and functional redundancy. *NPJ Syst Biol Appl* 4: 2

Cappelletti V, Hauser T, Piazza I, Pepelnjak M, Malinovska L, Fuhrer T, Li Y, Dörig C, Boersema P, Gillet L, *et al* (2021) Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* 184: 545–559.e22

Casado P, Rodriguez-Prados J-C, Cosulich SC, Guichard S, Vanhaesebroeck B, Joel S & Cutillas PR (2013) Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci Signal* 6: rs6

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD & Sander C (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* 39: D685–90

Chen B, Fan W, Liu J & Wu F-X (2014) Identifying protein complexes and functional modules--from static PPI networks to dynamic PPI networks. *Brief Bioinform* 15: 177–194

Chen L-L & Xiong Y (2020) Tumour metabolites hinder DNA repair. *Nature* doi:10.1038/d41586-020-01569-1 [PREPRINT]

Cocetta V, Ragazzi E & Montopoli M (2020) Links between cancer metabolism and cisplatin resistance. *Int Rev Cell Mol Biol* 354: 107–164

Drake JM, Paull EO, Graham NA, Lee JK, Smith BA, Titz B, Stoyanova T, Faltermeier CM, Uzunangelov V, Carlin DE, *et al* (2016) Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer. *Cell* 166: 1041–1054

Dugourd A, Kuppe C, Sciacovelli M, Gjerga E, Gabor A, Emdal KB, Vieira V, Bekker-Jensen DB, Kranz J, Bindels EMJ, *et al* (2021) Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol Syst Biol* 17: e9730

Dugourd A & Saez-Rodriguez J (2019a) Footprint-based functional analysis of multi-omic data. *Current Opinion in Systems Biology*

Dugourd A & Saez-Rodriguez J (2019b) Footprint-based functional analysis of multiomic data. *Current Opinion in Systems Biology* 15: 82–90

Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, Lerman JA, Lechner A, Sastry A, Bordbar A, *et al* (2016) Multi-omic data integration enables discovery of hidden biological regularities. *Nat Commun* 7: 13091

Emmanouilidi A & Falasca M (2017) Targeting PDK1 for Chemosensitization of Cancer Cells. *Cancers* 9

Engers R, Mrzyk S, Springer E, Fabbro D, Weissgerber G, Gernharz CD & Gabbert HE (2000) Protein kinase C in human renal cell carcinomas: role in invasion and differential isoenzyme expression. *Br J Cancer* 82: 1063–1069

Ewald JC, Matt T & Zamboni N (2013) The integrated response of primary metabolites to gene deletions and the environment. *Mol Biosyst* 9: 440–446

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, *et al* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 46: D649–D655

Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D & Saez-Rodriguez J (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* 29: 1363–1375

Garcia-Alonso L, Ibrahim MM, Turei D & Saez-Rodriguez J (2018a) Benchmark and integration of

resources for the estimation of human transcription factor activities. *bioRxiv*: 337915

Garcia-Alonso L, Iorio F, Matchan A, Fonseca N, Jaaks P, Peat G, Pignatelli M, Falcone F, Benes CH, Dunham I, *et al* (2018b) Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Res* 78: 769–780

Gaude E, Schmidt C, Gammage PA, Dugourd A, Blacker T, Chew SP, Saez-Rodriguez J, O'Neill JS, Szabadkai G, Minczuk M, *et al* (2018) NADH Shuttling Couples Cytosolic Reductive Carboxylation of Glutamine with Glycolysis in Cells with Mitochondrial Dysfunction. *Mol Cell* 69: 581–593.e7

Gleitz HFE, Dugourd AJF, Leimkühler NB, Snoeren IAM, Fuchs SNR, Menzel S, Ziegler S, Kröger N, Triviai I, Büsche G, *et al* (2020) Increased CXCL4 expression in hematopoietic cells links inflammation and progression of bone marrow fibrosis in MPN. *Blood* 136: 2051–2064

Gonçalves E, Raguz Nakic Z, Zampieri M, Wagih O, Ochoa D, Sauer U, Beltrao P & Saez-Rodriguez J (2017) Systematic Analysis of Transcriptional and Post-transcriptional Regulation of Metabolism in Yeast. *PLoS Comput Biol* 13: e1005297

Gonçalves E, Sciacovelli M, Costa ASH, Tran MGB, Johnson TI, Machado D, Frezza C & Saez-Rodriguez J (2018) Post-translational regulation of metabolism in fumarate hydratase deficient cancer cells. *Metab Eng* 45: 149–157

Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, *et al* (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* 46: D380–D386

Hernandez-Armenta C, Ochoa D, Gonçalves E, Saez-Rodriguez J & Beltrao P (2017) Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics* 33: 1845–1851

Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, Joughin BA, Stegle O, Lauffenburger DA, Heyn H, *et al* (2020) Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol* 21: 36

Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V & Sullivan M (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 40: D261–70

Horn H, Schoof EM, Kim J, Robin X, Miller ML, Diella F, Palma A, Cesareni G, Jensen LJ & Linding R (2014) KinomeXplorer: an integrated platform for kinome biology studies. *Nat Methods* 11: 603–604

Huang DW, Sherman BT & Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4: 44

Huan T, Palermo A, Ivanisevic J, Rinehart D, Edler D, Phommavongsay T, Benton HP, Guijas C, Domingo-Almenara X, Warth B, *et al* (2018) Autonomous Multimodal Metabolomics Data Integration for Comprehensive Pathway Analysis and Systems Biology. *Anal Chem* 90: 8396–8403

Huber W, von Heydebreck A, Sültmann H, Poustka A & Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18 Suppl 1: S96–104

Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Gonçalves E, Barthorpe S, Lightfoot H, *et al* (2016) A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* 166: 740–754

Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, Salic A, Asara JM, Lane WS & Kaelin WG Jr (2001) HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. *Science* 292: 464–468

Jaakkola P, Mole DR, -M. Tian Y, Wilson MI, Gielbert J, Gaskell SJ, v. Kriegsheim A, Hebestreit HF, Mukherji M, Schofield CJ, *et al* (2001) Targeting of HIF-alpha to the von Hippel-Lindau Ubiquitylation Complex by O2-Regulated Prolyl Hydroxylation. *Science* 292: 468–472 doi:10.1126/science.1059796 [PREPRINT]

Jelinek D & Wu X (2012) Faculty of 1000 evaluation for The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *F1000 - Post-publication peer review of the biomedical literature* doi:10.3410/f.14264142.15777309 [PREPRINT]

Jeske L, Placzek S, Schomburg I, Chang A & Schomburg D (2019) BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res* 47: D542–D549

Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27–30

Kawata K, Hatano A, Yugi K, Kubota H, Sano T, Fujii M, Tomizawa Y, Kokaji T, Tanaka KY, Uda S, *et al* (2018) Trans-omic Analysis Reveals Selective Responses to Induced and Basal Insulin across Signaling, Transcriptional, and Metabolic Networks. *iScience* 7: 212–229

Kharma B, Baba T, Matsumura N, Kang HS, Hamanishi J, Murakami R, McConechy MM, Leung S, Yamaguchi K, Hosoe Y, *et al* (2014) STAT1 drives tumor progression in serous papillary endometrial cancer. *Cancer Res* 74: 6519–6530

Kim S-Y & Volsky DJ (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6: 144

Kiparissides A & Hatzimanikatis V (2017) Thermodynamics-based Metabolite Sensitivity Analysis in metabolic networks. *Metab Eng* 39: 117–127

Köksal AS, Beck K, Cronin DR, McKenna A, Camp ND, Srivastava S, MacGilvray ME, Bodík R, Wolf-Yadlin A, Fraenkel E, *et al* (2018) Synthesizing Signaling Pathways from Temporal Phosphoproteomic Data. *Cell Rep* 24: 3607–3618

Krawczenko A, Bielawska-Pohl A, Wojtowicz K, Jura R, Paprocka M, Wojdat E, Kozłowska U, Klimczak A, Grillon C, Kieda C, *et al* (2017) Expression and activity of multidrug resistance proteins in mature endothelial cells and their precursors: A challenging correlation. *PLoS One* 12: e0172371

Krug K, Mertins P, Zhang B, Hornbeck P, Raju R, Ahmad R, Szucs M, Mundt F, Forestier D, Jane-Valbuena J, *et al* (2018) A curated resource for phosphosite-specific signature analysis. *Mol Cell Proteomics*

Kuehne A, Mayr U, Sévin DC, Claassen M & Zamboni N (2017) Metabolic network segmentation: A probabilistic graphical modeling approach to identify the sites and sequential order of metabolic regulation from non-targeted metabolomics data. *PLoS Comput Biol* 13: e1005577

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, *et al* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 44: W90–7

Lachmann A & Ma'ayan A (2009) KEA: kinase enrichment analysis. *Bioinformatics* 25: 684–686

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P & Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27: 1739–1740

Lim S, Lee S, Jung I, Rhee S & Kim S (2018) Comprehensive and critical evaluation of individualized

pathway activity measurement tools on pan-cancer data. *Brief Bioinform*

Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J & Saez-Rodriguez J (2019a) From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *NPJ Syst Biol Appl* 5: 40

Liu W, Payne SH, Ma S & Fenyö D (2019b) Extracting Pathway-level Signatures from Proteogenomic Data in Breast Cancer Using Independent Component Analysis. *Mol Cell Proteomics* 18: S169–S182

Lucas B, Grigo K, Erdmann S, Lausen J, Klein-Hitpass L & Ryffel GU (2005) HNF4 α reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma. *Oncogene* 24: 6418–6431

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, *et al* (2006) TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34: D108–D110

Maxwell PH, Wiesener MS, Chang GW, Clifford SC, Vaux EC, Cockman ME, Wykoff CC, Pugh CW, Maher ER & Ratcliffe PJ (1999) The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. *Nature* 399: 271–275

Melas IN, Sakellaropoulos T, Iorio F, Alexopoulos LG, Loh W-Y, Lauffenburger DA, Saez-Rodriguez J & Bai JPF (2015) Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. *Integr Biol* 7: 904–920

Miryala SK, Anbarasu A & Ramaiah S (2018) Discerning molecular interactions: A comprehensive review on biomolecular interaction databases and network analysis tools. *Gene* 642: 84–94

Network TCGAR & The Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499: 43–49 doi:10.1038/nature12222 [PREPRINT]

Osmanbeyoglu HU, Toska E, Chan C, Baselga J & Leslie CS (2017) Pancancer modelling predicts the context-specific impact of somatic mutations on transcriptional programs. *Nat Commun* 8: 14249

Pandey N, Lanke V & Vinod PK (2020) Network-based metabolic characterization of renal cell carcinoma. *Sci Rep* 10: 5955

Parikh JR, Klinger B, Xia Y, Marto JA & Blüthgen N (2010) Discovering causal signaling pathways through gene-expression patterns. *Nucleic Acids Res* 38: W109–17

Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D & Stuart JM (2013) Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 29: 2757–2764

Pirhaji L, Milani P, Leidl M, Curran T, Avila-Pacheco J, Clish CB, White FM, Saghatelian A & Fraenkel E (2016) Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nat Methods* 13: 770–776

Ramana CV, Grammatikakis N, Chernov M, Nguyen H, Goh KC, Williams BRG & Stark GR (2000) Regulation of c-myc expression by IFN-γ through Stat1-dependent and -independent pathways. *The EMBO Journal* 19: 263–272 doi:10.1093/emboj/19.2.263 [PREPRINT]

Reyna MA, Leiserson MDM & Raphael BJ (2018) Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* 34: i972–i980

Richard CL, Tan EY & Blay J (2006) Adenosine upregulates CXCR4 and enhances the proliferative and migratory responses of human carcinoma cells to CXCL12/SDF-1alpha. *Int J Cancer* 119:

2044–2053

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W & Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43: e47

Rodrigues P, Patel SA, Harewood L & Olan I (2018) NF-κB–Dependent Lymphoid Enhancer Co-option Promotes Renal Carcinoma Metastasis. *Cancer Discov*

Rolland-Turner M, Goretti E, Bousquenaud M, Léonard F, Nicolas C, Zhang L, Maskali F, Marie P-Y, Devaux Y & Wagner D (2013) Adenosine stimulates the migration of human endothelial progenitor cells. Role of CXCR4 and microRNA-150. *PLoS One* 8: e54135

Sayed-Ahmed MM (2010) Role of carnitine in cancer chemotherapy-induced multiple organ toxicity. *Saudi Pharm J* 18: 195–206

Schneider RK, Mullally A, Dugourd A, Peisker F, Hoogenboezem R, Van Strien PMH, Bindels EM, Heckl D, Büsche G, Fleck D, *et al* (2017) Gli1+Mesenchymal Stromal Cells Are a Key Driver of Bone Marrow Fibrosis and an Important Cellular Therapeutic Target. *Cell Stem Cell* 20: 785–800.e8

Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, Garnett MJ, Blüthgen N & Saez-Rodriguez J (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat Commun* 9: 20

Schwahn K & Nikoloski Z (2018) Data Reduction Approaches for Dissecting Transcriptional Effects on Metabolism. *Front Plant Sci* 9: 538

Sciacovelli M, Gonçalves E, Johnson TI, Zecchini VR, da Costa ASH, Gaude E, Drubbel AV, Theobald SJ, Abbo SR, Tran MGB, *et al* (2016) Fumarate is an epigenetic modifier that elicits epithelial-to-mesenchymal transition. *Nature* 537: 544–547

Sharifi-Noghabi H, Zolotareva O, Collins CC & Ester M (2019) MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 35: i501–i509

Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ & Cao K-AL (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 35: 3055–3062 doi:10.1093/bioinformatics/bty1054 [PREPRINT]

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, *et al* (2017) A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171: 1437–1452.e17

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550

Sulkowski PL, Oeck S, Dow J, Economos NG, Mirfakhraie L, Liu Y, Noronha K, Bao X, Li J, Shuch BM, *et al* (2020) Oncometabolites suppress DNA repair by disrupting local chromatin signalling. *Nature* doi:10.1038/s41586-020-2363-0 [PREPRINT]

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, *et al* (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43: D447–D452

Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P & Kuhn M (2016) STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 44: D380–D384

Tang A, Gao K, Chu L, Zhang R, Yang J & Zheng J (2017) Aurora kinases: novel therapy targets in

cancers. *Oncotarget* 8: 23937–23954

Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J & Frouin V (2014) Variable selection for generalized canonical correlation analysis. *Biostatistics* 15: 569–583

Tényi Á, de Atauri P, Gomez-Cabrero D, Cano I, Clarke K, Falciani F, Cascante M, Roca J & Maier D (2016) ChainRank, a chain prioritisation method for contextualisation of biological networks. *BMC Bioinformatics* 17: 17

Terfve CDA, Wilkes EH, Casado P, Cutillas PR & Saez-Rodriguez J (2015) Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat Commun* 6: 8033

Türei D, Korcsmáros T & Saez-Rodriguez J (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* 13: 966–967

Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, *et al* (2017) A pathology atlas of the human cancer transcriptome. *Science* 357

Välikangas T, Suomi T & Elo LL (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* 19: 1–11

Vanharanta S, Shu W, Brenet F, Hakimi AA, Heguy A, Viale A, Reuter VE, Hsieh JJ-D, Scandura JM & Massagué J (2013) Epigenetic expansion of VHL-HIF signal output drives multiorgan metastasis in renal cancer. *Nat Med* 19: 50–56

Vitrinel B, Koh HWL, Kar FM, Maity S, Rendleman J, Choi H & Vogel C (2019) Exploiting inter-data relationships in next-generation proteomics analysis. *Mol Cell Proteomics*

Wan B, Huang Y, Liu B, Lu L & Lv C (2019) AURKB: a promising biomarker in clear cell renal cell carcinoma. *PeerJ* 7: e7718

Wang K, Sun Y, Tao W, Fei X & Chang C (2017) Androgen receptor (AR) promotes clear cell renal cell carcinoma (ccRCC) migration and invasion via altering the circHIAT1/miR-195-5p/29a-3p/29c-3p/CDC42 signals. *Cancer Lett* 394: 1–12

Wiredja DD, Koyutürk M & Chance MR (2017) The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics*

Yang P, Patrick E, Humphrey SJ, Ghazanfar S, James DE, Jothi R & Yang JYH (2016) KinasePA: Phosphoproteomics data annotation using hypothesis driven kinase perturbation analysis. *Proteomics* 16: 1868–1871

Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, Jeon M, Kang J & Tan AC (2015) DSigDB: drug signatures database for gene set analysis. *Bioinformatics* 31: 3069–3071

Zhao H, Leppert JT & Peehl DM (2016) A Protective Role for Androgen Receptor in Clear Cell Renal Cell Carcinoma Based on Mining TCGA Data. *PLoS One* 11: e0146505

Zhou W-M, Wu G-L, Huang J, Li J-G, Hao C, He Q-M, Chen X-D, Wang G-X & Tu X-H (2019) Low expression of PDK1 inhibits renal cell carcinoma cell proliferation, migration, invasion and epithelial mesenchymal transition through inhibition of the PI3K-PDK1-Akt pathway. *Cellular Signalling* 56: 1–14 doi:10.1016/j.cellsig.2018.11.016 [PREPRINT]

Zhu G, Liang L, Li L, Dang Q, Song W, Yeh S, He D & Chang C (2014) The expression and evaluation of androgen receptor in human renal cell carcinoma. *Urology* 83: 510.e19–24

Zyla J, Marczyk M, Weiner J & Polanska J (2017) Ranking metrics in gene set enrichment analysis: do they matter? *BMC Bioinformatics* 18: 256