Modelling regulatory interactions between metabolism and signalling

Emanuel José Vieira Gonçalves

Hughes Hall

This dissertation is submitted on June, 2016 for the degree of Doctor of Philosophy
DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated.

The work presented in this dissertation was carried out at the EMBL-European Bioinformatics Institute under the supervision of Prof. Dr. Julio Saez-Rodriguez, between October 2012 and June 2016. This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified length limit of 60 000 words as defined by the Biology Degree Committee.

Emanuel José Vieira Gonçalves
June, 2016
Abstract

Cells are capable of adapting to a wide range of conditions and of performing many diverse biological tasks such as cell division, motility and differentiation. These are achieved by complex and coordinated intracellular responses involving gene regulatory, signalling transduction and metabolic processes. It is becoming apparent that these biological processes despite being biochemically distinct and having characteristic functions need to be considered as a whole to fully understand the functioning of a cell. With recent technological advances providing unprecedented coverage of the transcriptome, proteome and metabolome the bottleneck has shifted from data generation to the development of methods capable of integrating and interpreting these different types of biological data.

In the first two results chapters of my thesis I focus in yeast as a model organism. In Chapter 2, I introduce a novel mass-spectrometry based metabolomics data-set in yeast and pair it to existing phosphoproteomics measurements in the same conditions. I also demonstrate a framework to integrate these two data-sets to systematically hypothesise regulatory phosphorylation-sites in metabolic enzymes. In Chapter 3, I present a novel computationally efficient approach to estimate the activities of kinases and phosphatases using a compendium of published phosphoproteomics data-sets for which transcriptomics and metabolomics measurements are also available. I complement the assembled data-sets with the metabolomics and phosphoproteomics data-sets of Chapter 2. I also use the kinases/phosphatases approach to estimate the activity of transcription-factors using the transcriptomics data-sets. Consequently, I integrate the estimated protein activities with the metabolomics data-sets to systematically identify putative transcriptional and post-translational regulatory associations of metabolic processes.

In Chapters 4 and 5 I move from yeast to focus on more complex systems of cancer cell lines. In Chapter 4, I investigate the phosphoproteomics, proteomics
and metabolomics adaptations of renal cell cancer cell lines to the knockout of fumarate hydratase. I present a novel approach to identify potential regulatory kinases of metabolic pathways combining tailored approaches to model signalling transduction and metabolic processes. This allows me to hypothesise a mechanistic regulatory interaction between the metabolic enzyme pyruvate dehydrogenase and the protein tyrosine kinase ABL1. In the final results chapter of my thesis, Chapter 5, I analyse novel proteomics and phosphoproteomics data-sets acquired across a panel of colorectal cancer cell lines. I also show an integrative approach to systematically find functional and direct protein interactions, and then expand it to identify hypothetic genomic mutations or post-translational modifications that alter the protein abundance and possibly impair protein-protein interactions.
Acknowledgements

During the past four years I have been supported by many people and this work benefited directly and indirectly from their help, for that I am deeply grateful.

First and foremost, a special thanks goes to my supervisor, Julio Saez-Rodriguez, for providing me this fantastic opportunity, for trusting in me even when I would have not done myself, and for always lending a helping hand. This would have not been possible without him.

From all the people that I had the pleasure to work with I would like to reserve a special thanks to Pedro Beltrão for his mentorship and for always having a spare minute, more like half-an-hour, for my scientific questions. Also, many thanks to my TAC committee, Christoph Steinbeck, Kiran Patil and Christian Frezza, for their supervision and for ensuring that I was on the right track.

I am very thankful to Marco Sciacovelli, Sofia Costa, Edoardo Gaude, Isaac Johnson, Vinny Zecchini and their group leader Christian Frezza for always welcoming me in their lab and for a rewarding collaboration. I would like to extend this thanks to David Ochoa, Paolo Casale and Claudia Hernandez for very useful and fruitful discussions. I also thank Uwe Sauer and all his group for hosting me in their lab for a short-term internship and for providing me a truly insightful experience. In particular, I would like to thank Zrinka Nakic, Mattia Zampieri, Markus Basan, Dimitris Christodoulou and Ruben Mars for their friendship and support. I am also grateful to Daniel Machado, Sara Correia and Miguel Rocha for perpetuating a fruitful scientific collaboration with my previous research group in Braga, Portugal. I also would like to thank Theodoros Roumeliotis, Steven Williams, Ultan McDermott and Jyoti Choudhary for many insightful scientific discussions and collaboration.

A big thanks to all the present and past members of the Saez-Rodriguez group at EBI and in Aachen for all their support and for making the everyday life at the lab an enjoyable, happy and rewarding experience. Specifically, I would
like to thank Thomas Cokelaer, Marti Bernardo, Federica Eduati, Jakob Wirbel, Luz Garcia, Francesco Iorio, Michael Schubert, Michael Menden, Camille Terfve, Damien Arnol, Fatemeh Zamanzad, Luis Tobalina, Claudia Hernandez, Dénes Türei, Aurélien Dugourd, Angeliki Kalamara and Luca Cerone with whom I have directly worked or benefited from their support. But these are only some of the members of our group and everyone was important in its own way.

I am thankful to the EMBL, in particular EBI, predoctoral community for providing a stimulating and helpful environment, which shaped me professionally and personally.

This experience did not start only four years ago, in fact it would not have been possible without my parents, Isabel Vieira and Sérgio Gonçalves, my sister and brother, Sofia Gonçalves and Sérgio Gonçalves, and my closest family, Maria João Verdasca, Tomás Gonçalves and Simão Gonçalves. To them I am very grateful for every bit of support, counseling and for always looking after my best interests.

Lastly, but surely not least, a warm thank you to Joana for being by my side on another journey, for always providing me sensible advice and, more than that, for everything else.
CONTENTS

Abstract v
Acknowledgements vii

Contents viii

List of Figures xiii

Glossary xx

1 Introduction 21
  1.1 Complex cell phenotype ........................................ 21
  1.2 Principles and modelling approaches of cell signalling .... 28
  1.3 Principles and modelling of cell metabolism ................. 35
  1.4 Modelling the integration of signalling and metabolism .... 40
  1.5 Aims of the thesis ............................................. 44

2 Identification of functional phosphorylation sites in metabolic enzymes using time-course phosphoproteomics and metabolomics 47
  2.1 Introduction ................................................... 47
  2.2 Results ....................................................... 50
    2.2.1 Characterisation of metabolomics changes upon salt and
          pheromone stimulation ...................................... 50
    2.2.2 Phosphoproteomics profiles of osmotic and pheromone re-
          sponse .......................................................... 53
    2.2.3 Metabolic similarity between stimulation and co-stimulation 56
    2.2.4 Systematic identification of regulatory phosphosites in meta-
          bolic enzymes ............................................... 57
3 Systematic analysis of transcriptional and post-transcriptional regulation of metabolism in yeast

3.1 Introduction

3.2 Results

3.2.1 Compendium of yeast metabolic, transcriptional and phosphoproteomic paired data-sets

3.2.2 Inferring enzymatic activity of kinases, phosphatases and transcription-factors

3.2.3 Growth rate implications in intracellular changes

3.2.4 Estimating metabolic changes from transcription-factors, kinases and phosphatases activities

3.2.5 Inferring putative regulatory protein-metabolites interactions

3.3 Discussion

4 Signalling regulation of metabolism in hereditary and leiomyomatosis renal cell cancer

4.1 Introduction

4.2 Results

4.2.1 Molecular characterisation of HLRCC cell lines

4.2.2 Genome-scale modelling of metabolism

4.2.3 Systematic analysis of associations between phosphorylation and flux changes

4.2.4 Identification of kinases/phosphatases involved in the regulation of metabolic enzymes

4.3 Discussion

5 Systematic characterisation of the proteomics and phosphoproteomics interactome in colorectal cancer

5.1 Introduction

5.2 Results

5.2.1 Experimental setup and data normalisation

5.2.2 Proteome and phosphoproteome of colorectal cancer cell lines

5.2.3 Estimating the activity of kinases and phosphatases
5.2.4 Systematic identification of protein-protein interactions and protein complexes .................................................. 112
5.2.5 Impact of mutations and phosphorylation in protein-protein associations ............................................................... 114
5.2.6 Finding proteomics based predictions of drug response .......... 118
5.3 Discussion ........................................................................................................................................ 118

6 Conclusions ...................................................................................................................................... 123

A Computational methods and materials ................................................................................................. 129
A.1 Code dependencies and availability ........................................................................................................ 129
A.2 Regulatory networks .............................................................................................................................. 130
  A.2.1 Yeast kinase/phosphatase network ........................................................................................................ 130
  A.2.2 Yeast transcription-factor regulatory network .................................................................................... 130
  A.2.3 Human kinase/phosphatase network .................................................................................................. 131
A.3 Linear regression models ....................................................................................................................... 131
  A.3.1 Identification of putative regulatory phosphosites in metabolic enzymes .............................................. 131
  A.3.2 Estimating metabolomics changes using transcription-factor and kinases/phosphatases activities .......... 132
  A.3.3 Identifying regulatory interactions between kinases, phosphatases and transcription-factors and metabolites ...... 133
  A.3.4 Normalise phosphorylation with protein abundance ....................................................................... 134
  A.3.5 Protein-pairs correlations and associations with protein modifications .................................................. 134
  A.3.6 Drug response associations with proteomics, phosphoproteomics and protein activities ......................... 135
A.4 Sampling intracellular metabolic flux distributions .................................................................................. 136

B Experimental methods and materials ........................................................................................................ 137
B.1 Yeast cells .............................................................................................................................................. 137
  B.1.1 Strain, growth and sample preparation ............................................................................................... 137
  B.1.2 Acquisition of intracellular metabolite levels and quality control ....................................................... 138
  B.1.3 Compendium of high throughput publicly available yeast studies .................................................. 139
B.2 Human cells
B.2.1 Hereditary leiomyomatosis and renal cell cancer cell lines
  B.2.1.1 Growth and sample preparation
  B.2.1.2 Consumption and release quantification of metabolites
  B.2.1.3 Mass-spectrometry based proteomics and phosphoproteomics
B.2.2 Colorectal cancer cell lines
  B.2.2.1 Cell culture and reagents
  B.2.2.2 Protein digestion and TMT labeling
  B.2.2.3 Peptide fractionation
  B.2.2.4 Phosphopeptide enrichment
  B.2.2.5 LC-MS Analysis
  B.2.2.6 Protein identification and quantification

C Supplementary figures
C.1 Additional results for Chapter 2
  C.1.1 Comparison of untargeted metabolomics normalisation to 0 seconds and 25 seconds
  C.1.2 Time profiles of representative ions of the untargeted metabolomics with a fold-change compared to time 0 seconds and 25 seconds
  C.1.3 Overlap of significant phosphorylation changes among different conditions
  C.1.4 Metabolite connectivity association with measured phosphosites
C.2 Additional results for Chapter 3
  C.2.1 Agreement between GSEA and Linear regression to estimate enzymatic activities
  C.2.2 Principal component analysis of data-sets and correlation with relative growth rate
  C.2.3 Top predicted metabolites
  C.2.4 ROC-curve analysis of the average feature coefficients
C.3 Additional results for Chapter 4
C.3.1 Replicates distributions of the consumption/release metabolomics experiment 156
C.3.2 Hierarchical clustering of the replicates of the consumption/release experiment 157
C.3.3 Correlation between proteomics and transcriptomics 158
C.3.4 Unsupervised clustering of the proteomics biological replicates 159
C.3.5 Hierarchical clustering of the phosphoproteomics replicates 160
C.3.6 Kinases/phosphatases-substrates associations histogram 161
C.4 Additional results for Chapter 5 161
C.4.1 Proteomics and phosphoproteomics batch effect 162
C.4.2 Spearman correlation distributions between phosphosites and their protein abundance 163
C.4.3 Unsupervised hierarchical clustering of phosphoproteomics measurements not corrected by protein abundance 164
C.4.4 HNF4A association with the proteome 165
C.4.5 Proteomics pairwise spearman correlations 166
C.4.6 Drug response associations with phosphoproteomics and kinases and phosphatases activity 167

Bibliography 169
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Osmotic stress response in yeast</td>
<td>23</td>
</tr>
<tr>
<td>1.2</td>
<td>Schematic comparison of energy production between Warburg effect and oxidative phosphorylation.</td>
<td>25</td>
</tr>
<tr>
<td>1.3</td>
<td>Cellular regulatory interactions</td>
<td>27</td>
</tr>
<tr>
<td>1.4</td>
<td>Principles of signalling pathways</td>
<td>31</td>
</tr>
<tr>
<td>1.5</td>
<td>Mass-spectrometry experimental procedures</td>
<td>35</td>
</tr>
<tr>
<td>1.6</td>
<td>Metabolic pathways and regulatory processes</td>
<td>36</td>
</tr>
<tr>
<td>1.7</td>
<td>Regulatory interactions between signalling and metabolism</td>
<td>42</td>
</tr>
<tr>
<td>1.8</td>
<td>Thesis content word cloud</td>
<td>45</td>
</tr>
<tr>
<td>2.1</td>
<td>Regulation of glycerol accumulation upon osmotic stress</td>
<td>48</td>
</tr>
<tr>
<td>2.2</td>
<td>Overview of the experimental design</td>
<td>52</td>
</tr>
<tr>
<td>2.3</td>
<td>Phosphorylation changes of the different conditions</td>
<td>55</td>
</tr>
<tr>
<td>2.4</td>
<td>Analysis of the association between each stimulation and the co-stimulation.</td>
<td>57</td>
</tr>
<tr>
<td>2.5</td>
<td>Workflow to identify putative regulatory phosphorylation-sites in metabolic enzymes</td>
<td>58</td>
</tr>
<tr>
<td>2.6</td>
<td>Results of the linear regression models</td>
<td>59</td>
</tr>
<tr>
<td>3.1</td>
<td>Analysis and experimental design and data consistency.</td>
<td>68</td>
</tr>
<tr>
<td>3.2</td>
<td>Protein activity analysis</td>
<td>72</td>
</tr>
<tr>
<td>3.3</td>
<td>Computational framework for predicting metabolic adaptation with TFs and kinases/phosphatases activities</td>
<td>76</td>
</tr>
<tr>
<td>3.4</td>
<td>Overview of the putative protein-metabolite regulatory interactions</td>
<td>78</td>
</tr>
<tr>
<td>4.1</td>
<td>Diagram depicting the different biological measurements acquired and used to model signalling and metabolism in HLRCC cell lines</td>
<td>86</td>
</tr>
</tbody>
</table>
4.2 Biological measurements acquired for HLRCC cell lines ............................ 88
4.3 Genome-scale metabolic models constrained using the measured consumption and release rates of metabolites ................................. 92
4.4 Associations between phosphosites in metabolic enzymes and their catalysed reactions ................................................................. 94
4.5 Kinases and phosphatases activities ......................................................... 95
4.6 Kinase/phosphatase-substrate interaction network representation ............ 96
4.7 Possible regulatory mechanism of PDHA1 S232 by ABL1 kinase ............... 98

5.1 Experimental overview and reproducibility analysis of the acquired data-sets ................................................................. 107
5.2 Proteomics and phosphoproteomics cell lines correlation profiles ......... 109
5.3 Kinases activity scores for KRAS and BRAF and downstream targets 111
5.4 Inferring protein-protein interactions from proteomics measurements 113
5.5 Associations between mutations and protein-pairs ............................... 115
5.6 Representative significant associations between protein-protein residuals and protein modifications .......................................... 117
5.7 Significant drug response associations with proteomics measurements 119

C.1 Comparison of untargeted metabolomics normalisation to 0 seconds and 25 seconds ................................................................. 148
C.2 Time profiles of representative ions of the untargeted metabolomics with a fold-change compared to time 0 seconds and 25 seconds .... 149
C.3 Overlap of significant phosphorylation changes among different conditions .................................................................................. 150
C.4 Metabolite connectivity association with measured phosphosites .......... 151
C.5 Agreement between GSEA and Linear regression to estimate enzymatic activities .................................................................................. 152
C.6 Principal component analysis of of data-sets and correlation with relative growth rate ........................................................................... 153
C.7 Top predicted metabolites ...................................................................... 154
C.8 ROC-curve analysis of the average feature coefficients ....................... 155
C.9 Replicates distributions of the consumption/release metabolomics experiment .................................................................................. 156
C.10 Hierarchical clustering of the replicates of the consumption/release experiment .............................................. 157
C.11 Correlation between proteomics and transcriptomics .............. 158
C.12 Unsupervised clustering of the proteomics biological replicates . 159
C.13 Hierarchical clustering of the phosphoproteomics replicates .... 160
C.14 Kinases-phosphatases/substrates associations histogram .......... 161
C.15 Proteomics and phosphoproteomics batch effect .................. 162
C.16 Spearman correlation distributions between phosphosites and their protein abundance ........................................... 163
C.17 Unsupervised hierarchical clustering of phosphoproteomics measurements not corrected by protein abundance ............ 164
C.18 HNF4A association with the proteome ............................. 165
C.19 Proteomics pairwise spearman correlations ........................ 166
C.20 Drug response associations with phosphoproteomics and kinases and phosphatases activity ............................... 167
Glossary


ABL1 tyrosine-protein kinase ABL1.

AMPK AMP-activated protein kinase.

CCM central carbon metabolism.

CORE consumption release experiment.

EMT epithelialmesenchymal transition.

FBA flux balance analysis.

FH fumarate hydratase.

Gpd1 glycerol-3-phosphate dehydrogenase.

GS gas chromatography.

GSEA gene-set enrichment analysis.

GWAS Genome-wide association studies.

HLRCC hereditary leiomyomatosis and renal cell cancer.

IDH isocitrate dehydrogenase.

IMAC immobilized metal affinity chromatography.

LC liquid chromatography.
MAPK  mitogen-activated protein kinases.

MCMC  Markov Chain Monte Carlo.

MOAC  metal oxide affinity chromatography.

MS  mass-spectrometry.

OD  optical density.

ODE  ordinary differential equation.

OLS  Ordinary least squares.

PC  principal component.

PCA  principal component analysis.

PDH  pyruvate dehydrogenase.

PSM  peptide spectra mass.

PTM  post-translation modification.

QTOF  quadrupole time-of-flight.

RNA  ribonucleic acid.

SH2  Src homology 2.

SNP  single nucleotide point mutation.

TCA  tricarboxylic acid.

TFs  transcription-factors.

TMT  tandem mass tag.
INTRODUCTION

1.1 Complex cell phenotype

Through the process of evolution cells have acquired intricate systems that coordinate a vast amount of different tasks, to name a few, cell differentiation, cell motility and cell growth [Alberts et al., 2008]. Understanding the molecular events underpinning these cellular adaptations has proven to be a remarkable and insightful challenge. Cell responses are mediated by a variety of biochemical molecules each showing particular functional characteristics that make them important for specific biological roles. For example, DNA is responsible for storing the genomic material of the cell, metabolites are important for sustaining the cellular homeostasis and proteins are involved in the catalysis of a multitude of reactions. Nevertheless, despite their differences and the fact that they are involved in different biological processes these molecules are all dependent on each other and as a whole they regulate the phenotype of the cell [Hartwell et al., 1999]. Biological processes are composed, among others, by gene, protein, metabolic and miRNA networks [Davidson and Erwin, 2006; Hartwell et al., 1999]. In contrast to prokaryotic, eukaryotic cells also display a complex organisation of the molecular space with membrane defined compartments, such as nucleus, mitochondria or golgi apparatus [Alberts et al., 2008]. This segregation of the cellular space into compartments allows cells to perform, for instance, compartment-specific reactions and adjust the catalytic activity of proteins by localising them to specific compartments. The complete extent of the organisation
of cellular processes is still not yet fully understood. In this thesis, I focus on two model organisms: yeast and human. Yeast is a widely used organism and has been extensively characterised to study signalling transduction and metabolism, and human cells are of general relevance with many potential applications in the field of biomedicine. Moreover, the underlying principles and analysis methods can potentially be translated between the two, for example, some of the metabolic phenotypes in yeast can also be seen in human cells [Vander Heiden et al., 2009].

Cellular adaptations require a coordinated response of the different biological processes and this interplay is an ubiquitous phenomenon across all organisms. An illustrative example of the importance of the interplay among these processes for the cell phenotype is the osmotic stress in the single-celled microorganism *Saccharomyces cerevisiae* [Saito and Posas, 2012; Hohmann, 2002; Dove et al., 1997; Hinnebusch and Johnston, 2011] (Figure 1.1). Upon extracellular osmotic imbalance, stress sensors in the cell membrane are activated and trigger the intracellular response by using an evolutionarily conserved signal transduction pathway, mitogen-activated protein kinases (MAPK), via protein phosphorylation [Jacoby et al., 1997; Kültz and Burg, 1998]. Protein kinase Hog1 [Posas et al., 1996] has a key role mediating this intracellular response and double phosphorylation in specific residues renders the protein kinase fully active after one minute of sensing the imbalance [Reiser et al., 1999; Kanshin et al., 2015]. Its activation initiates quick functional changes in the activity of ion membrane transporters and increases the catalytic activity of metabolic enzymes involved in the production of glycerol [Saito and Posas, 2012]. Hog1 can also translocate to the nucleus and modulate more permanent responses by regulating the activity of transcription factors (TFs) and consequently adapt the expression of broad sets of genes and regulate cell cycle progression [Saito and Posas, 2012]. Yeast is a relatively simple organism and osmotic imbalance is a common stress that these organisms need to face. In spite of that, this response already displays the necessity of the integrated response using several biological processes, i.e. signalling is responsible for sensing the extracellular imbalance and initiate the intracellular response to coordinate quick changes in the abundance of specific metabolites and adapt the expression of large sets of genes.

The intricate cellular coordination displayed in yeast is also visible in human cells, and its deregulation can have harmful implications in the stability of the cells and of the whole organism manifesting as diseases, including, cystic fibrosis,
**Figure 1.1:** Osmotic stress response in yeast. Stress sensors located in the cell membrane detect the osmotic imbalance and initiate an intracellular response by altering the phosphorylation status of several proteins. Hog1 is the central signalling protein in this response and it controls, among others, membrane transporters, metabolic processes and translation. Hog1 can also translocate to the nucleus where it can affect the expression of several genes and cell cycle progression.

diabetes and cancer [Ramsey et al., 2011; Mootha et al., 2003; Davies et al., 1994; Kelley et al., 2002; Davies et al., 2002; Weinberg, 2007]. Cancer is largely driven by somatic mutations where cells undergo an abnormal accumulation of mutations in the genome [Stratton et al., 2009; Martincorena and Campbell, 2015]. Some of these mutations harbour particular functional alterations that lead cells to behave very distinctively from their original phenotype [Weinberg, 2007]. While cancer cells display very heterogenous mutation landscapes they retain ubiquitous phenotypes, termed the hallmarks of cancer [Hanahan and Weinberg, 2000, 2011]. To name a few, cancer cells avoid growth suppressors and sustain increased growth with constitutively active signalling pathways and display an abnormal metabolic phenotype with deregulated energy production [Hsu and Sabatini, 2008; Pavlova and Thompson, 2016; Lynch et al., 2004; Vivanco and Sawyers, 2002]. Specific types of mutations can constitutively activate protein kinases, namely, BRAF
single nucleotide point mutation (SNP) within the kinase domain in position 600 alters a valine to a glutamic acid (V600E) [Davies et al., 2002; Wan et al., 2004], and this is frequently seen in different cancer types [Forbes et al., 2015]. BRAF kinase phosphorylates downstream MAPK pathway and contributes to the deregulation of cellular growth. Another almost ubiquitous phenotype of cancer cells is the increased glucose consumption and increased lactate secretion [Warburg et al., 1924; Warburg, 1956] (Figure 1.2). Cancer cells display a similar behaviour as fast proliferating cells, they opt for aerobic glycolysis regardless the presence or absence of oxygen [Warburg, 1956; Vander Heiden et al., 2009]. From the point of view of efficiency on the total amount of energy that can be produced from a molecule of glucose this is a puzzling and still elusive phenomenon [Pfeiffer et al., 2001; Pfeiffer and Bonhoeffer, 2002]. Diverting glucose consumption to the production of lactate only generates approximately 4 mol ATP/mol glucose whereas the full respiration of glucose through the tricarboxylic acid cycle TCA cycle in the mitochondria generates approximately 36 mol ATP / mol glucose [Warburg et al., 1924; Warburg, 1956; Vander Heiden et al., 2009] (Figure 1.2). This seemingly wasteful processes is usually referred to as the Warburg effect. Warburg initially suggested that an impaired mitochondrial function would be the reason for the diversion of pyruvate into lactate [Cairns et al., 2011]. Although, more recently alternative hypothesis supports that aerobic glycolysis is a consequence of cancer cells minimizing the biosynthesis of proteins [Basan et al., 2015; Hsu and Sabatini, 2008]. Indeed the enzymatic cost of producing energy through aerobic glycolysis is lower than using oxidative phosphorylation which requires the whole mitochondrial enzymatic machinery. Recently, a study carried in Escherichia coli quantified this difference and estimated that fermentation is 50% more efficient than respiration for energy production [Basan et al., 2015]. The Warburg effect being a characteristic and almost ubiquitous phenomenon in cancer cells is one of the most tempting therapeutical windows in cancer metabolism [Galluzzi et al., 2013]. Cancer hallmarks emphasise the importance of the regulatory interactions between biological processes and that a deregulation in one can lead to a system-wide imbalance modifying the cell state and alter the original healthy phenotype [Tan et al., 2009; Vidal et al., 2011; Menche et al., 2015].

The interaction between the different biological processes is achieved by an extensive and diverse network of regulatory interactions [Vidal et al., 2011]. Along the central dogma of biology many types of regulatory associations occur be-
Figure 1.2: Diagram recapitulating the metabolic differences in the central carbon metabolism of the Warburg effect, aerobic glycolysis, compared to the oxidative phosphorylation. Aerobic glycolysis is the preferred way of tumours to produce energy where pyruvate is diverted from the TCA cycle to the production of lactate, generating approximately 4 mol ATP/mol of glucose. Whereas, if the full respiratory chain is used a total of approximately 36 mol of ATP can be produced per mol of glucose.

between distinct biochemical entities (Figure 1.3). As shown in the osmotic stress response in yeast, phosphorylation can regulate the activity of metabolic enzymes and consequently alter metabolism [Saito and Posas, 2012] (Figure 1.2, 1.3 A). Some studies have identified the importance of specific phosphorylation-sites to regulate the activity of metabolic enzymes involved in the central carbon metabolism (CCM) [Oliveira et al., 2012; Chassagnole et al., 2002]. Protein phosphorylation is only one of the many possible post-translation modifications (PTMs),
e.g. acetylation, methylation, succinylation and glycosylation [Hornbeck et al., 2015]. Acetylation has broad relevance in the regulation of the proteome, in particular, metabolic enzymes [Choudhary and Mann, 2010; Weinert et al., 2014; Choudhary et al., 2014; Hunter, 2007b] (Figure 1.3 B). PTMs besides their specific role they are known to crosstalk and in combination regulate the proteome [Hunter, 2007b; Zhao et al., 2014; Venne et al., 2014]. For example, phosphodegron events [Hunter, 2007b] are phosphorylation changes that mark proteins for ubiquitination and consequently degradation. Some studies have developed approaches to study these interactions in a high-throughput manner [Swaney et al., 2013; Mertins et al., 2013] (Figure 1.3 C). Genomic mutations of protein kinases is of utmost importance in the field of cancer biology where a significant percentage of the most mutated genes are protein kinases and therefore these are potential drug targets [Fleuren et al., 2016; Zhang et al., 2009; Cohen, 2002b] (Figure 1.3 D). Mutations can also affect the catalytic activity of metabolic enzymes and thereby alter metabolism. Another regulatory interaction between different biochemical processes is the epigenetic control of gene expression by modifying the structure of DNA by adding methylation markers. This is a well known regulatory event very important in tissue development and specification [Bird, 2007] and also important in cancer [Jones and Laird, 1999; Esteller, 2008] (Figure 1.3 E). Recent studies have shown that patients with brain gliomas harboring mutations in the metabolic enzyme isocitrate dehydrogenase (IDH) [Choi et al., 2012], global epigenetic changes can be mediated by an abnormal accumulation of the metabolite 2-hydroxyglutarate [Xu et al., 2011], and thereby coining the term oncometabolite. Also, in Hereditary Leiomyomatosis and Renal Cell Cancer (HLRCC) patients, mutations in the metabolic enzyme Fumarate Hydratase (FH) disrupt the catalytic function of the protein and leads to increased levels of its reactant, fumarate [Frezza et al., 2011]. Intracellular accumulation of fumarate can then trigger different signalling cascades, namely, HIF pathway [Isaacs et al., 2005]. Many other types of interactions between different biochemical processes of the cell could be specified, for example, miRNAs regulation (Figure 1.3 F) and protein allosteric regulation (Figure 1.3 G). The regulatory processes enumerated before emphasise the complexity and the diversity of the biochemical interactions occurring within a cell.

The complexity of biological processes made apparent that accurate and scalable representations of the biological networks were needed, thus the interdis-
Figure 1.3: Diagram depicting some of the existing regulatory interactions between different molecules in the cell. A) Phosphorylation regulation of proteins and their role in signaling cascades. B) Acetylation regulation of proteins. C) Cross-talk between post-translation modifications, where phosphorylation can mark for ubiquitination. D) Gain-of-function mutation leading to increased protein activity. E) Gene expression regulated by methylation in the promoter region of the gene. F) Transcript expression regulation by miRNA binding and consequently degradation. G) Protein allosteric regulation by binding of small molecules. CpG methylation islands are marked by circle sticks where filled circles represent methylated CpG sites and empty ones represent demethylated sites.
disciplinary field of systems biology emerged [Kitano, 2002; Hartwell et al., 1999]. Systems biology developed approaches that can range from coarse-grained models, that capture the dynamic behaviours of molecules, to large-scale networks, that comprise genome-scale perspectives of the cellular adaptations [Kreeger and Lauffenburger, 2010; Ideker and Krogan, 2012; Palsson, 2006; Klipp et al., 2011]. On the next sections of this chapter I will enumerate and describe several modelling approaches developed to represent and analyse signalling and metabolic networks, and more recent methods to integrate both.

1.2 Principles and modelling approaches of cell signalling

Reversible phosphorylation of proteins is a broad regulatory mechanism with implications in nearly all processes of the cell. The understanding of phosphorylation as a regulatory mechanism started with the work of Carl and Gerty Cori [Cori and Green, 1943], who discovered that the enzyme catalyzing the conversion of glycogen to glucose-1-phosphate was present in two differently active forms, phosphorylase a and b. Although, they incorrectly argued that the conversion from one form to the other was achieved by allosterically regulation of the nucleotide 5’AMP [Cori and Green, 1943]. The true regulatory mechanism behind the interconversion between phosphorylase a to b was latter shown to be a phosphorylation/dephosphorylation event [Fischer and Krebs, 1955; Krebs and Fischer, 1955; Sutherland and Wosilait, 1955]. In particular, Fischer and Krebs also showed that the conversion of form b to a could be achieved in the presence of Mg-ATP and an enzyme, which they called phosphorylase kinase [Fischer and Krebs, 1955; Krebs and Fischer, 1955]. The importance of this mechanism was later emphasised by the discovery that incubating liver tissue with adrenalin or glucagon lead to increased glycogenolysis and increased phosphorylase a activity [Sutherland, 1971]. Thus directly showing that the enzyme catalysing the rate-limiting step of glycogenolysis, phosphorylase enzyme, was regulated by an hormone response. Several years later it was discovered that another metabolic enzyme, pyruvate dehydrogenase (PDH), could also be regulated by phosphorylation, in particular it could be inactivated by phosphorylation [Linn et al., 1969]. These different evidences together with many other studies [Corbin et al., 1970; Hunter and Sefton,
1980; Sadowski et al., 1986] strengthen the perspective of phosphorylation as a
mechanism of general importance in regulating cellular processes. Nowadays, it
is expected that 30% of the total proteome covalently bind a phosphate group
[Cohen, 2002a; Lim and Pawson, 2010]. Several aspects of phosphorylation, such
as its simplicity, reversibility and the availability of ATP as a phosphoryl donor,
helps to understand this importance as a general regulatory mechanism [Cohen,
2002a]. Phosphorylation is only one of the many different types of PTMs, nev-
nevertheless its by far the most well studied modification with the highest number
of studies reported in current databases [Lu et al., 2013]. This existing bias in
the literature can be mainly attributed to the availability of specific and sensi-
tive reagents that could be reliably used to identify phosphorylation events. For
example, reversible acetylation of lysines was discovered only a decade later than
phosphorylation [Allfrey et al., 1964], nonetheless due to technical challenges in
the identification of acetylation sites it was only much later that it was possible
to identify the first non-histone acetylation target [L'Hernault and Rosenbaum,
1985].

Protein function can be modulated by the covalent addition of a phosphate
group to one or more amino-acids. This modification occurs in specific residues
and these are, generally, serines (S), tyrosines (Y) and threonines (T). A phos-
phate group carries two negative charges, therefore when added to the protein it
can cause conformational changes to it by, for example, attract positively charged
regions of the protein [Seet et al., 2006]. These conformational changes can have
strong implications in the activity of the protein as it can promote or impede
the binding of ligands to the protein surface. As phosphorylation is a reversible
reaction, removing the phosphate group restores the original conformation of the
protein. Moreover, the addition of a phosphate group can create binding struc-
tures that other proteins can recognise and attach to. Protein domains, such as
Src homology 2 (SH2) [Sadowski et al., 1986], bind to certain peptides sequences
with a phosphorylated residue. Consequently, protein phosphorylation play an
important role in protein-protein interaction, in particular it can modulate the
assembly of protein complexes. As a consequence phosphorylation controls the
activity, structure and cellular localisation of proteins involved in a wide range of
biological processes.

Phosphorylation is enzymatically-catalysed by protein kinases and phospho-
tases which add or remove a phosphate group, respectively. Besides its functional
impact in protein function, phosphorylation is also important for sensing a variety
of extracellular cues and to propagate its signal intracellularly in what is often
viewed as linear cascades of kinases [Seger and Krebs, 1995; Chang and Karin,
2001] (Figure 1.4 A). This phenomena of signalling pathways is commonly termed
as signalling cascade [Chang and Karin, 2001]. Kinases have affinity to specific
protein sequences motifs and thereby target and bind specific phosphorylation-
sites where they catalyse the phosphorylation reaction [Manning et al., 2002a].
The specificity of the kinases allows to classify them into two major classes: a
broader group that targets S and T residues, and another family that targets Y
residues [Alberts et al., 2008; Manning et al., 2002a]. Within these superfamilies
smaller sub-groups can be defined where protein kinases display specificity to
particular regions surrounding the target residue, flanking regions. The impor-
tance of the flanking regions of the phosphorylation-sites to construe the possible
regulatory kinase is well accepted but the exact size of the region is still debat-
able [Obenauer et al., 2003]. Computational methods have successfully predicted
novel targets of kinases using their previously reported substrates and the flank-
ing regions of the phosphosites [Obenauer et al., 2003; Miller et al., 2008; Linding
et al., 2007]. Many other approaches have since then been presented to model ki-
nase specificity considering also other types of information, such as, proximity of
the kinase and the substrate in protein-protein interaction networks [Horn et al.,
2014; Wagih et al., 2016; Schwartz et al., 2009; Huang et al., 2005].

There are approximately 500,000 potential phosphosites in the human phos-
phoproteome [Lemeer and Heck, 2009] and only a small fraction of these are as-
associated with the approximately 500 kinases identified in human [Manning et al.,
2002b; Hornbeck et al., 2015]. Considering that phosphorylation changes are
mainly driven by kinases or phosphatases, this supports the idea that only a
small portion of the phosphorylation interactome has been characterised. This
limited characterisation can be due to the transient nature of kinase/phosphatase-
substrate interactions and thereby the difficulty to capture them. Phosphopro-
teomics measurements display an added level of complexity since only a small
percentage of the identified phosphosites have been functionally annotated [Bel-
trao et al., 2012]. In fact, considering that phosphorylation residues are generally
poorly conserved across species it was hypothesised that a large percentage of
the currently identified phosphosites have no functional impact [Landry et al.,
2009; Beltrao et al., 2009; Oliveira et al., 2012]. Moreover, often multiple phos-
Figure 1.4: Representation of principles of signalling pathways. A) Membrane receptors sense extracellular cues and generate an internal phosphorylation response. B) Multiple phosphosites may be required to regulate protein activity. C) Signalling proteins may display self-phosphorylation events. D) Positive or negative feedback loops in signalling pathways. E) Cross-talk between different signalling networks.

phosphosites are required to be (de)phosphorylated in order to display a functional impact in the protein (Figure 1.4 B). For instance, Hog1 kinase is only fully active when adjacent T and Y residues, T174 and Y176, are phosphorylated [Kanshin et al., 2015]. In contrast, decreased phosphorylation in S24 and S27 in Glycerol-3-phosphate dehydrogenase (Gpd1) renders the metabolic enzyme active [Oliveira et al., 2012]. Once active a kinase may induce self-phosphorylation events that are required to achieve full enzymatic activity (Figure 1.4 C). Phosphoproteomics measurements provide a snapshot of the proteins undergoing phosphorylation changes, nevertheless to obtain functional information low throughput assays are required or other functional readouts need to be integrated [Beltrao et al., 2013].

In order to attain functional information from the phosphoproteome snapshot some methods have been proposed to estimate the activity of kinases by assessing the changes on their reported substrates [Casado et al., 2013; Mischnik et al., 2016; Ochoa et al., 2016]. Knowledge about kinase/phosphatase-substrate interactions have been accumulating in multiple resources [Hornbeck et al., 2015; Dinkel et al., 2011; Sadowski et al., 2013; Perfetto et al., 2016; Gnad et al., 2011,
Methods to estimate kinases activities use lists of reported kinases-substrates interactions to convey information about the activity of the regulator. If the measured targets of a kinase are in general decreasing, or increasing, in phosphorylation this represents that the kinase activity has decreased, or increased, respectively. This approach can be applied to different conditions and provides a step forward into deriving functional information from phosphoproteomics. The fact that a kinase activity is generally estimated relying on several phosphosites decreases the sparseness of the measurements and thereby increases the overlap among conditions. The current major limitations of these methods are the incomplete knowledge and complexity of the kinases interactome [Jørgensen and Linding, 2010] and the often narrow overlap between the measured phosphorylation sites and those reported in literature [Hornbeck et al., 2015]. Therefore, these limit the robust quantitative estimation of activities to a few dozens or hundreds of kinases.

Phosphoproteomics data-sets provide rich resources to identify cellular processes undergoing regulation, however to understand the dynamic flow of the signal prior knowledge is required. Mathematical models can be used to represent and model the kinetics of signalling pathways [Downward, 2001; Barrios-Rodiles et al., 2005; Kholodenko, 2006; MacNamara et al., 2012]. Signalling networks display complex dynamic behaviours exerted by, for example, negative and positive feedback loops (Figure 1.4 D) and crosstalk between different pathways (Figure 1.4 E). This allows cells to regulate the intensity and the frequency of the signal response and it is a vital characteristic for multiple cell responses [Kholodenko, 2006; Mitchell et al., 2015]. The dynamic behaviour of signalling pathways can be well represented using ordinary differential equations (ODEs) [Gonzalez et al., 2006; Aldridge et al., 2006]. ODE models provide accurate representations of the dynamic phosphorylation status of a protein in a signal response, although it requires prior knowledge on the signalling network and also information on the enzymatic rates of the reactions [Chang et al., 2009; Milo et al., 2010]. Thus, the application of these models are restricted by the availability of these type of information and the complexity of the model makes it computationally intractable for large networks. More scalable modelling approaches of signalling networks resorted to simpler representations, for example, boolean logic [Saez-Rodriguez et al., 2009; Klamt et al., 2006]. Boolean logic can be used to represent the activity status of the protein as either active or inactive, and model
the interactions among proteins as an activating or inhibiting effect with logic and or operations. Boolean approaches to integrate experimental data and infer context-specific networks have been proposed [Fauré et al., 2006]. Extensions to this approach include an expanded representation of the boolean status of the protein to a continuous range with Fuzzy-logic [Aldridge et al., 2009], integration of boolean time-courses [MacNamara et al., 2012] and time courses representation were expanded to continuous ranges with logic based ODE models [Wittmann et al., 2009; Krumsiek et al., 2010]. Although, these type of analysis to achieve robust results require data-sets with several perturbations along the same pathway and are still highly dependent on the prior information about the pathway. Considering that, to circumvent this limitation a few studies presented methods to integrate phosphoproteomics and other omics data-sets with large protein-protein interaction networks [Huang and Fraenkel, 2009; Tuncbag et al., 2012; Huang et al., 2013]. These approaches have resorted to an efficient algorithm [Dittrich et al., 2008] for solving a prize collecting steiner tree problem. A steiner tree solution provides a minimum-spanning tree that converges proteins that are changing significantly in the biological conditions at hand and are associated in the protein-protein interaction network. These methods model phosphoproteomics data in large-scale networks, although they overlook an important complex aspect of signalling proteins, that kinases and phosphatases target specific phosphosites and therefore protein-protein networks are an inadequate representation. A recent approach resorting to boolean formalism integrates large-scale signalling networks and phosphoproteomics data to extract context-specific kinase/phosphatase-substrate networks [Terfve et al., 2015]. This method provides signalling networks that maximise the agreement between the prior knowledge and the experimental data. Deconvoluting the signalling transduction from phosphoproteomics measurements is still a challenge since much of the interactome is still missing and a large part of the modifications do not have an apparent functional implication.

Several experimental methods, including, mass-spectrometry (MS) based approaches and reverse phase protein microarrays have been developed to measure phosphorylation changes [Aebersold and Mann, 2003; Domon and Aebersold, 2010; Mann et al., 2002; Sheehan et al., 2005]. Mass-spectrometry is the method of choice to acquire large-scale proteomics and phosphoproteomics data-sets [Aebersold and Mann, 2003; Olsen and Mann, 2013]. To measure protein abundance
one starts by extracting the protein material from the samples and consequently perform a protease step to cleave proteins into smaller peptides. Peptides are then injected into a chromatograph and then measured in the MS (Figure 1.5 A). Previously to enter the MS the sample mixture needs to be physically separated using chromatography technologies, most commonly liquid-chromatography (LC) or gas-chromatography (GS) (Figure 1.5 A). Phosphoproteomics measurements also require a phosphopeptide enrichment step due to their naturally lower abundance and lack of sensitivity of the MS instruments [Zhou et al., 2013; Beltran and Cutillas, 2012]. Two of the most broadly used approaches are immobilized metal affinity chromatography (IMAC) [Tape et al., 2014] and metal oxide affinity chromatography (MOAC) [Fila and Honys, 2012; Beltran and Cutillas, 2012]. A variety of different MS setups and configurations can be used and these will influence the portion of the proteome and phosphoproteome that is measured [Riley and Coon, 2016]. Exploratory MS approaches are generally divided into two groups: label-free and isotope labeling. Label-free quantification approaches have simpler experimental procedures as they do not require any chemical modification of the peptides and can measure a theoretically unlimited number of samples as they do not rely on multiplexed systems. Nevertheless, due to the stochastic nature of the ion selection in each MS run different portions of the proteome are measured in each sample and thereby this greatly affects the overlap among samples [Liu et al., 2004; Nesvizhskii, 2007; Worboys et al., 2014]. Isotope labeling approaches can partially address this issue, in particular an increasingly used approach is tandem mass tags (TMT)-MS [McAlister et al., 2012; Rauniyar and Yates, 2014]. Currently, TMT-MS supports up to ten multiplexed systems by labeling the peptides with ten different isobaric mass tags. Multiplexed systems allow to measure several samples in a single MS run and therefore avoiding the sparseness problem of the stochastic coverage of the proteome across different MS runs. After isobaric labeling the peptides go through a first MS run which provides a spectral mass for each peptide, then a second MS run is necessary to distinguish the isobaric tags of each peptide, thus generating a unique peptide spectra mass (PSM) for each multiplexed sample. TMT-MS is a powerful approach to robustly measure the proteome of different samples, nevertheless the number of samples that can be plexed is limited and interbatch comparison will suffer from the same limitation of label-free MS. Besides, both label-free and labelled approaches are generally biased towards highly abundant peptides
as these are more frequently detected and reliably measured. These approaches enable the measurement of thousands of peptides and phosphopeptides, nevertheless still only a small fragment of the total theoretical phosphoproteome can be captured in a single experiment [Lemeer and Heck, 2009].

![Diagram of proteomics and TMT-MS](image)

**Figure 1.5:** Illustration of the necessary steps for a mass-spectrometry based proteomics analysis (A) and in particular for a TMT-MS experiment (B). (A) Protein extraction and digestion are the two initial steps of a MS experiment, these are then generally followed by chromatography and by MS measurement. (B) For TMT the samples are labelled with isobaric tags and pulled together before entering in the chromatography. Quantification of the peptide or phosphopeptide abundance is then performed with two MS, MS/MS, where the first is used to quantify the different peptides and the second one to identify the isobaric label of each sample. Phosphoproteomics analysis require an extra phosphorylation enrichment step.

### 1.3 Principles and modelling of cell metabolism

Cell metabolism is constituted by two generally inverse processes: catabolic reactions which break down molecules for the production of energy, and anabolic reactions which provide the essential blocks to synthesise molecules, such as proteins, using energy produced by catabolic reactions [Alberts et al., 2008]. Metabolism is a vital part of the cell and is responsible for producing energy, maintaining cell homeostasis and mediating cellular adaptations to environmental nutritional changes. These functions are achieved by taking nutrients into the cell, for in-
stance amino acids, and consequently use them as substrates to sustain the activity of given metabolic pathways. These pathways are generally regulated by enzymes that catalyse specific metabolic reactions that would normally not occur, or occur at much lower rates [Seibert and Tracy, 2014] (Figure 1.6 A). Hence, metabolic enzymes have an important role in the maintenance and regulation of metabolism and have proven to be insightful in understanding how cells adjust their phenotype [Palsson, 2006, 2015]. Metabolic reactions catalysed by enzymes often form linear pathways that are responsible for the production of metabolites or cofactors necessary for the cell, for example, glycolysis pathway converts glucose into pyruvate through several consecutive metabolic reactions and produces energy in the form of ATP [Alberts et al., 2008].

![Figure 1.6: Illustration of metabolic pathways and some of their regulatory mechanisms. A) Many metabolic reactions are catalysed by metabolic enzymes. B) Metabolites can directly regulate metabolic enzymes by allosteric regulation. C) Metabolic reactions are often reversible and the abundance of the product and substrate may saturate its conversion rate. D) Metabolic reactions can also be irreversible. Cofactors are represented with small filled circles.](image)

Metabolic adaptations are generally manifested by alterations in metabolites abundance and changes in reaction rates. These display dynamic and complex profiles that are driven by self-regulatory metabolic mechanisms [Gruning et al., 2010], including, allosteric inhibition of enzymes by metabolites (Figure 1.6 B)
and reaction flux control by the imbalance between substrate and product concentration (Figure 1.6 C). This dynamic behaviour can be accurately modelled using ODE models [Palsson, 2011; Chassagnole et al., 2002; Costa and Vinga, 2016; Link et al., 2013; Noguchi et al., 2013]. Nevertheless, the same limitations of ODE models for signaling networks apply here, there is a limited knowledge of the enzymatic rates in different conditions and dynamic models do not scale to large networks. Thus, dynamic modelling is limited to central carbon metabolism and a few diverging pathways [Miskovic et al., 2015; Link et al., 2014; Chandra et al., 2011]. Kinetic models have proven to be valuable tools to understand different metabolic regulatory processes, such as, allosteric protein-metabolite interactions [Link et al., 2013; König et al., 2012].

Despite the fact that metabolism is a highly dynamic process, if a cell is growing in a stable environment it will stabilise its response and will reach a steady-state. In a steady-state the cell achieves a mass-balance by keeping the total concentration of metabolites and reaction rates constant. This important observation together with stoichiometric information of the metabolic reactions enabled to model metabolism with a simple linear mathematical formulation, termed Flux Balance Analysis (FBA) [Varma and Palsson, 1994; Orth et al., 2000]. This formulation represents each metabolite as a linear equation:

\[
\frac{dm}{dt} = R_1 + 2R_2 - R_3 \quad (1.1)
\]

where the metabolite concentration over time is dependent of the reactions that produces, R1 and R2, and consumes it, R3, taking in consideration the stoichiometric requirements of each reaction. This equation is further simplified since FBA relies on the mass-balance assumption:

\[
0 = R_1 + 2R_2 - R_3 \quad (1.2)
\]

Therefore, considering that there is no change in the metabolite abundance the reactions that consume and produce the metabolite will cancel out. Further constraints are added to the model to represent thermodynamic constraints of the reactions:

\[
0 < R_1 < \infty \quad (1.3)
\]
This, enables to represent reversible (Figure 1.6 C) and irreversible (Figure 1.6 D) reactions. Equations (1.2) to (1.4) are used to represent all the possible fluxes of the reactions in the metabolic network and these represent the solution space of the model. This space can then be searched by using specific objective functions, one of the most widely used is the maximisation of the biomass production, which mimics the main goal of microorganisms of maximising growth [Lewis et al., 2012; O’Brien et al., 2015]. FBA was firstly introduced to model the metabolic network of microorganisms and obtained robust estimations of growth rate and intracellular flux rates [Savinell and Palsson, 1992; Varma and Palsson, 1994; Orth et al., 2000; Palsson, 2006]. In contrast to ODE models, FBA provided for the first time a scalable and flexible representation to model metabolic networks that could be expanded to a genome-scale level. Developments in genomics allowed to sequence entire genomes of organisms and this laid the foothold for the reconstruction of genome-scale metabolic models for several organisms [Price et al., 2004; Monk et al., 2014; Palsson, 2015]. A variety of other methods expanding the capabilities and applications of FBA have been proposed [Lewis et al., 2012]. For example, parsimonious-FBA (pFBA) represents a parsimonious usage of the metabolic pathways where the maximisation of the biomass is then followed by the minimisation of all the metabolic fluxes. This finds a solution that guarantees the maximum possible biomass but also minimises the usage of metabolic enzymes [Lewis et al., 2010]. Genome-scale models are also very useful tools to assess the implication of metabolic enzymes knockouts and several methods have been presented for identifying optimal combinations of gene deletions and over/under expression that optimises the production of compounds of interest [Burgard et al., 2003; Rocha et al., 2008; Gonçalves et al., 2012; Rocha et al., 2010]. FBA based approaches and genome-scale models are established methods to model metabolism [O’Brien et al., 2015; Bordbar et al., 2014; Monk et al., 2014] and these have been used in biotechnological and biomedical applications [Patil et al., 2005; Zomorrodi et al., 2012; Rocha et al., 2010; Shlomi et al., 2011; Frezza et al., 2011].

The sequencing of the human genome opened the possibility of building the first human genome-scale metabolic model [Duarte et al., 2007; Thiele et al.,
Modelling human metabolism possess additional challenges than microorganisms, namely, human cells do not generally strive for maximising growth, tissue cells have very specific roles for which the metabolic objective is not known or easily represented [Jerby et al., 2010; Agren et al., 2012; Jerby et al., 2012]. Moreover, contrary to microorganisms growing media, human cells are grown in rich media for which the exact constituents are not defined. Albeit, for cancer cells the biomass reaction can arguably still be considered since tumour cells lose their tissue of origin role and acquire abnormal growth rates [Folger et al., 2011; Yizhak et al., 2015]. Promising results have been shown in the field of cancer metabolism using genome-scale models [Yizhak et al., 2015] to explore synthetic lethality [Frezza et al., 2011], identify drug targets that revert metabolic states [Yizhak et al., 2013; Folger et al., 2011] and provide insights into the possible protein synthesis regulation leading to the Warburg effect [Shlomi et al., 2011]. Human metabolic reconstructions have also been broadly used to generate tissue-specific metabolic models accounting for the genetic imprint and proteome expression [Gatto et al., 2014; Mardinoglu et al., 2014; Agren et al., 2014, 2012; Jerby et al., 2010; Correia and Rocha, 2015]. Accurate modelling of the human metabolism still poses many challenges, for instance, the complexity of the network and the arguably more complex regulatory mechanisms present in mammalian cells.

Similarly to signalling, MS is the method of choice to measure metabolic changes. MS based approaches can measure from dozens to hundreds of metabolites intra and extracellularly in many different conditions [Dettmer et al., 2007; Villas-Bôas et al., 2005; Jain et al., 2012; Hakimi et al., 2016; Fuhrer et al., 2011]. Metabolite measurements can be performed in a targeted way where a previously defined set of metabolites are compared to a carbon labeled reference. This provides technically robust measurements but the possible set of metabolites to analyse is limited. In contrast, an exploratory analysis can be performed with direct infusion untargeted metabolomics where metabolite samples are analysed directly with MS and the spectra are then matched against a reference library [Fuhrer et al., 2011]. This approach provides higher coverage of the metabolome but comes at the expense of an increased uncertainty on the annotation, since metabolites with the same mass are indistinguishable [Fuhrer et al., 2011]. While these two distinct approaches provide a snapshot of the metabolome more dynamic measurements can be obtained by using carbon labelled metabolic substrates [Zamboni et al., 2009; Sauer, 2006; Fischer et al., 2004; Wiechert,
2001]. Labelling substrates, such as glucose or glutamine, with carbon-13 (13C) isotopes followed by intracellular metabolomics allows to measure the number of labeled carbon isotopes in each metabolite and thereby identify the metabolic pathways that use the labelled substrates [Christen and Sauer, 2011; Sauer, 2006; Büscher et al., 2009]. This approach is particularly useful to understand how the intracellular fluxes are organized in branching points of the metabolic network, e.g. branching between glycolysis and pentose-phosphate pathway [Fischer et al., 2004]. The main limitation of carbon labelled metabolomics is that it is restricted to a set of well known pathways for which the carbon changes in the metabolic reactions are known, these are necessary to correctly track the labeled substrate. Another frequent type of metabolic measurements are the quantification of consumption and release rates of metabolites to the medium. Metabolites concentration in the media are measured after a fixed interval of time of adding the cells, and then compared to the fresh media baseline metabolomics measurements. Subsequently, quantitative exchange rates can be calculated using the time difference and an approximate cell count. Quantitative assessment of the rates is important as it discards possible general effects of cell size and provides measurements adequate for using genome-scale metabolic models [Jain et al., 2012; Tedeschi et al., 2013]. These measurements are often complemented with intracellular carbon labelling data to quantify intracellular fluxes of metabolic pathways [Fischer et al., 2004; Zamboni et al., 2009]. Different biological experiments can be performed to measure the adaptation of metabolic pathways, nevertheless due to the requirement of prior knowledge they are generally focused in well characterised pathways, such as, CCM.

1.4 Modelling the integration of signalling and metabolism

Metabolism and signalling govern two very distinct types biological processes. Still, it is clear that they do not operate alone (Figure 1.7) and that representing them separately overlooks a vast part of biological events that affect the cellular phenotype [Chubukov et al., 2014; Saltiel and Kahn, 2001]. On the one hand, signalling can directly shape metabolism, for instance, by regulating the activity of metabolic enzymes via phosphorylation [Oliveira et al., 2012, 2015b; Hu et al.,
On the other hand, signalling can indirectly impact metabolism, for example, by altering the activity of TFs that regulate the expression of metabolic enzymes [Gerosa et al., 2013; Melas et al., 2011] (Figure 1.7 B). While phosphorylation is arguably the most well studied other PTMs, such as acetylation, have important implications in metabolic processes [Weinert et al., 2014; Choudhary et al., 2014; Schölz et al., 2015] (Figure 1.7 C). Similarly to phosphorylation, acetylation is enzymatically catalysed and reversible [Delcuve et al., 2012]. Membrane transporters can also be actively regulated by signalling pathways and thereby control the exchange of metabolites, this particular example is seen in the osmotic stress response in yeast with the regulation of ion transporters [Saito and Posas, 2012] (Figure 1.7 D). Metabolites can regulate allosterically protein activity, for example, of TFs and thereby have a broad implication in gene expression [Oliveira et al., 2015a] (Figure 1.7 E) and of protein kinases [Chaneton et al., 2012; Grüning et al., 2011] (Figure 1.7 F). Additionally, it is important to consider that the time scales of these processes may be different [Lee et al., 2008]. For example, signalling transduction phosphorylation changes can occur in a short interval, for instance, Hog1 in yeast shows a complete activation after one minute [Kanshin et al., 2015] and insulin responses in mouse harbor significant changes in AKT1 after 30 seconds [Humphrey et al., 2015]. Real-time metabolome profiling shows that, similarly to signalling, metabolic processes have fast adaptations to environmental changes [Link et al., 2015]. In contrast, gene expression regulation displays longer response times as it requires transcription and translation to occur to have an impact in metabolism [Ralser et al., 2009; Zaslaver et al., 2004; Oliveira et al., 2015a] (Figure 1.7 G).

Conceptually, signalling and metabolic pathways are also apart, for example, signalling networks represent a flow of signal, while metabolic networks represent exchange rates. Hence, these two biological processes were classically modelled separately and specific mathematical representations and approaches were developed [Gonçalves et al., 2013; Feist et al., 2009; Papin et al., 2005; Hyduke and Palsson, 2010]. Increased availability of multiple types of data across the same conditions opened the possibility to systematically explore the interactions between these processes and characterise functional implications among them [Yugi et al., 2014; Schulz et al., 2014; Bodenmiller et al., 2010; Hakimi et al., 2016; Oliveira et al., 2015b; Jain et al., 2012]. The adaptation of the phosphoproteome in yeast to different growing conditions have been characterised and
Figure 1.7: Diagram representing possible regulatory interactions between signalling and metabolic networks. A) Metabolic enzymes can be directly regulated by phosphorylation events catalysed by kinases or phosphatases. B) Signalling proteins can, by phosphorylation, regulate the activity of transcription-factors that in turn can regulate the expression of metabolic enzymes. C) Acetylation may regulate metabolic enzymes. D) Signalling pathways can control membrane transporters and thereby regulate the exchange of metabolites and cofactors between the medium and the cell. E, F) Metabolites can allosterically bind and regulate proteins, e.g. enzymes, kinases and transcription-factors. G) Enzymes expression can regulate metabolic reactions.

explored in combination with metabolic flux estimations to identify functionally relevant phosphosites in metabolic enzymes [Oliveira et al., 2012]. In a follow up study, Oliveira et al. [Oliveira et al., 2015b,a] acquired phosphoproteomics, metabolomics and transcriptomics in yeast upon three different nutrient perturbations to explore the regulatory implication of signalling in metabolism [Oliveira et al., 2015b] and also the metabolic mediated regulation of TORC1-dependent transcriptome [Oliveira et al., 2015a]. A recent studied presented a novel approach to integrate time-course resolved phosphoproteomics and metabolomics measurements to characterise the signal flow response to acute stimulation with insulin and hypothesise regulatory interactions between the two biological processes [Yugi et al., 2014]. These studies illustrate the importance of integrated analysis of both types of biological measurements to identify regulatory mecha-
nisms between signalling and metabolism.

Integrating signalling and metabolism can borrow concepts from different approaches used to integrate other types of omics, for example, the integration of transcriptional regulation and metabolism [Gerosa et al., 2015; Patil and Nielsen, 2005; Zeleznik et al., 2014; Gerosa et al., 2013; Covert et al., 2004]. A recent study showed how steady-state metabolomics and transcriptomics measurements can be used to identify the key transcriptional regulators mediating metabolic adaptations of *E. coli* under multiple growing conditions [Gerosa et al., 2015]. Also, genome-scale models have been widely used to explore transcriptional regulation in metabolic enzymes and estimate their functional impact by using FBA-based approaches. A variety of methods have been proposed to estimate flux changes from metabolic enzymes expression levels and have been applied to several organisms including human [Shlomi et al., 2008; Becker and Palsson, 2008; Schmidt et al., 2013; van Berlo et al., 2011; Colijn et al., 2009; Brandes et al., 2012; Jensen and Papin, 2011]. A thorough benchmark of these methods interestingly showed that gene expression brings little predictive power over a standard pFBA, a method that does not account for any other type of biological data and simply minimises the total flux in the metabolic model [Machado and Herrgård, 2014]. This important benchmark emphasises the complexity of integrating different types of biological measurements. Some conceptual methods have been proposed to integrate signaling, metabolism and gene regulation processes using genome-scale reconstructions, i.e. integrated FBA (iFBA) [Covert et al., 2008] and integrated dynamic FBA (idFBA) [Lee et al., 2008]. iFBA presented an approach to integrate a gene-regulatory boolean network and a simple ODE model of signalling phosphotransferase in a genome-scale metabolic model of *E. coli* [Covert et al., 2008]. Another integrated approach, termed idFBA, was presented in yeast showing an integrated stoichiometric reconstruction of the three biological processes, incorporating slow and fast reactions in the framework [Lee et al., 2008]. Slow reactions are incorporated directly into the stoichiometric matrix with a time-delay, while fast reactions rely on the pseudo steady-state assumption of FBA. Recently, a new generation of genome-scale models have been presented, ME-models, that expand previous reconstructions and allow to simulate metabolism together with macromolecular expression by constraining the mathematical formulation with transcriptional, translational and enzyme rates constraints [O’Brien et al., 2013; Lerman et al., 2012; Thiele et al., 2012]. These
models showed increased accuracy compared to previous metabolic reconstructions but are only available for a few of organisms, computationally expensive and require complex algorithms to simulate. Classic ODE models can accurately model both signalling and metabolism providing insights into the regulatory interactions between the two [Mosca et al., 2012; König et al., 2012]. However, this is limited to well characterised interactions for which several parameters are known, and does not scale to large networks, as mentioned before.

Ultimately, integrative approaches aim to comprehensively represent all the biological modules and their diverse types of interactions to robustly predict the cellular phenotype. The model that came closer to this was the whole-cell model of Mycoplasma genitalium [Karr et al., 2012]. Expanding on the basis of iFBA [Covert et al., 2008] this approach handles each of the considered biological modules separately, representing them with the most adequate mathematical formalism and using tailored methods to simulate them. For example, FBA was used to model metabolism whereas ribonucleic acid (RNA) and protein degradation were modelled as Poisson processes. Simulations of the whole-model are performed in discrete time steps using the assumption that each module is independent in short timescales. Therefore, each biological module is simulated autonomously and the shared variables are then updated and used by all the dependent modules in the next step simulation. This faithfully shows the challenges and the complexity that one needs to take in consideration to be able to model an organism as a whole system, even for such simple unicellular organism with only approximately 500 genes reported. While this concept work paves the way in how this can be achieved it is fair to assume that we are still far from such models for more complex organisms, such as E. coli, yeast and in particular human cells.

1.5 Aims of the thesis

As outlined above, the classical independent perspective of signalling and metabolism is being rapidly replaced by comprehensive and integrated panels of thousands of measurements characterising both biological processes. This gave rise to many unanswered questions and the need for methods that are computationally efficient and statistically robust. Therefore, my work has explored and contributed to current challenges with novel studies in the following way:
• developed efficient mathematical approaches to acquire functional information from high-throughput phosphoproteomics, metabolomics and transcriptomics data-sets using large-scale interaction networks;

• built integrated analysis workflows that explore associations between protein regulators activity profiles and metabolic adaptations in steady-state and in time course resolved experiments;

• explored direct functional implications of signalling in metabolism by identifying regulatory phosphorylation sites in metabolic enzymes;

• acquired time-resolved intracellular metabolomics measurements that together with other novel data-sets, acquired by collaborators, were used to develop the methods presented in this work;

Figure 1.8: Word cloud of the whole thesis content.

All the results chapters of this thesis are devoted to the integration of phosphoproteomics measurements with different omics with particular emphasis on the integration with metabolism (Figure 1.8). Chapters 2 and 3 are based on novel and published experimental data-sets measuring metabolomics and phosphoproteomics adaptation of the yeast *Saccharomyces cerevisiae* to a variety of steady-state and dynamic perturbations. In particular, Chapter 2 introduces a novel metabolomics data-set of yeast under different environmental conditions that I
acquired. It also presents an approach to integrate phosphoproteomics and metabo-
holomics to explore functional regulatory phosphosites at a genome-scale level. Chapter 3, shows a computationally efficient approach to estimate kinase and phosphatases activities from high-throughput phosphoproteomics and explores regulatory functional kinases/phosphatases-metabolic associations. Chapters 4 and 5 focus on cancer using human cell lines as biological models. Specifically, Chapter 4 presents novel proteomics, phosphoproteomics and metabolomics data-
sets characterising the adaptation of hereditary leiomyomatosis and renal cell can-
cer cell lines to the reconstitution of the metabolic enzyme fumarate hydratase. I also present a novel and systematic approach to infer signalling regulators of metabolic enzymes using tailored approaches to model signalling and metabolism. Lastly, in Chapter 5 I investigate novel two data-sets, proteomics and phospho-
proteomics, acquired across a broad panel of colorectal cell lines and I present an approach to evaluate the implication of mutations and PTMs in protein abundance and protein-protein interactions.
2.1 Introduction

Signal transduction is a vital response mechanism for cells to adapt their phenotype to extracellular cues. Protein phosphorylation in the yeast *Saccharomyces cerevisiae* is thought to control a large percentage of the total proteome and be involved in the regulation of many basic biological functions [Holt et al., 2009; Taylor et al., 2012]. Yeast has been extensively used as a model organism to
study the functional implications of phosphorylation changes. Current technological advances have allowed to systematically and comprehensively measure the phosphoproteome across dozens of genetic modifications [Bodenmiller et al., 2010], environmental adaptations [Oliveira et al., 2012] and intracellular signals [Daub et al., 2008]. Nonetheless, analysing phosphoproteomics data is still a challenging task since for many of the phosphorylation changes measured the functional impact is not known or relevant [Beltrao et al., 2012] and therefore direct association with biological insight is hard [Oliveira et al., 2012].

Metabolic processes can be directly regulated by signalling networks by phosphorylation of enzymes [Oliveira et al., 2012]. With the development of metabolomics experimental methods [Schulz et al., 2014] and modelling approaches [Mo et al., 2009] it is now possible to model the metabolomic phenotype at a genome-scale and thereby estimate the activity of the metabolic enzymes. Enzymatic activity changes can then be paired with phosphorylation changes and provide associations between phosphorylation changes and metabolic activity and thereby possible functional information of phosphorylation-sites [Oliveira et al., 2012].

Figure 2.1: Upon osmotic imbalance stress sensors located in the membrane of yeast cells are activated. This leads to the activation of a phosphorylation signalling response where Hog1 is the key regulatory protein. Hog1 is responsible for mediating several cellular adaptations, in particular in metabolism it controls the expression and phosphorylation of enzymes involved in the production of glycerol, which is then retained in the cell.
In this chapter, I explored possible regulatory phosphorylation-sites of metabolic enzymes by designing and performing time-resolved metabolomics experiments under salt and pheromone stimuli, pairing an existing phosphoproteomics experiment [Vaga et al., 2014]. Both salt and pheromone are known to promote changes in phosphorylation of MAPK pathway and specifically they share Ste11, Ste20 and Cdc24 protein kinases. To study the crosstalk the authors [Vaga et al., 2014] measured the initial phosphoproteomics changes up to 45 minutes for each condition separately and for all the possible time-wise combinations. Hog1 kinase is at the core of the response to osmotic stress and is responsible to trigger many cellular adaptations such as controlling ion transports, glycerol production and cell cycle [Hohmann, 2002; Petelenz-Kurdziel et al., 2013; Saito and Posas, 2012]. Upon osmotic stress Hog1 regulatory phosphorylation-sites, T174 and Y176, are phosphorylated and thereby increase the protein catalytic activity. Additionally, the main metabolic response to salt stress is the production and retention of glycerol to balance the osmotic pressure [Saito and Posas, 2012]. Glycerol biosynthesis is regulated by metabolic enzymes, such as Glycerol-3-phosphate dehydrogenase (Gpd1/2) and Glycerol-3-phosphate phosphatase (Gpp1/2), via changes in phosphorylation. Decreased phosphorylation in S24 and S27 of Gpd1 is associated with increased activity of the metabolic enzymes and consequently with increased production of glycerol-3-phosphate (Figure 2.1). The mechanistic implications of osmotic stress in metabolism in yeast is better understood than the pheromone induced alterations. Yeast cells mate in the presence of favourable conditions and the opposite mating type [Merlini et al., 2013; Bardwell, 2005]. In these conditions yeast cells release pheromones that once sensed initiate cell migration and an intracellular response mediating cell cycle arrest and cell wall conformation changes [Merlini et al., 2013]. While phenomenologically the impact of pheromone response has been well characterized the metabolic specific response is still poorly characterised.

Here, I present a novel dynamic metabolomics data-set capturing the response of yeast cells to salt and pheromone perturbations. The metabolite measurements are combined with existing phosphorylation measurements in the same conditions [Vaga et al., 2014] to explore possible regulatory interactions between signalling and metabolism. In particular, I present a computational approach to identify putative regulatory phosphosites of metabolic enzymes that are responsible for mediating the metabolic adaptation. This approach identifies possible regula-
tory phosphorylation sites in Gpd1 responsible for the production of glycerol 3-phosphate and sites in the trehalose 6-phosphate synthase/phosphatase complex (Tsl1) important for mediating the production of trehalose.

2.2 Results

2.2.1 Characterisation of metabolomics changes upon salt and pheromone stimulation

To explore regulatory phosphorylation changes in metabolic enzymes I set to pair the phosphoproteomics [Vaga et al., 2014] with intracellular metabolic measurements (Figure 2.2 A). To simplify the experimental design co-stimulation was only performed for same time exposure for both perturbations and thereby only three different conditions were measured, i.e. salt, pheromone and salt plus pheromone (see section B.1) (Figure 2.2 B). From the phosphoproteomics experiments it was reported that cell cultures stimulated with pheromone exhibited a long period before initiating the signalling response, while salt stimulation displayed an almost immediate response [Vaga et al., 2014]. Hence, to have both perturbations occurring at similar time-scales Cdc28 analog sensitive strains were used and were treated with 10 μM of Cdc28 ATP inhibitor one hour before stimulation. Experiments were performed in biological triplicates with cells inoculated in shake flasks and Cdc28 inhibition synchronized at and optical density (OD) of 0.6. After one hour of being exposed to Cdc28 inhibitor the initial time-points, 0 minutes, of each condition were extracted and cells were filtered with fast-filtration and then suspended with cold-extraction. Then 0.4 M of NaCl was added to the first shake flask, 1uM of pheromone to the second and 0.4M of NaCl and 1uM of pheromone to the third. Samples using the same extraction method as before were collected at higher time resolution than the phosphoproteomics data-set, 0 and 25 seconds and 1, 4, 5, 9, 10, 15, 20, 25, 35 and 45 minutes. This allows to get a better characterisation of the metabolomics changes, nevertheless to integrate with the phosphoproteomics data-set only the overlapping time-points were used. Metabolites were measured using two different MS approaches: (i) an internal 13C labelled standard was added to the samples and were analysed using LC-MS/MS [Buescher et al., 2010], termed targeted metabolomics, (ii) and an
exploratory approach [Fuhrer et al., 2011] was performed by direct flow double injection of extracts on a quadrupole TOF MS (QTOF-MS), termed untargeted metabolomics (Figure 2.2 C). The second approach allows higher coverage at the expense that metabolites with the same mass are indistinguishable. The untargeted metabolomics samples were run with double injection of the same samples and these were considered as technical duplicates. Three biological replicates were acquired for each experiment in separate days.

Targeted metabolomics allowed the identification of 54 metabolites, 26 of which were kept after quality control filtering based on the metabolite spectra (see section B.1.2). Samples were log2 transformed and each metabolite was normalised to the 13C internal standard, then fold-changes were calculated comparing to the initial time-point, 0 seconds, as in the phosphoproteomics study. Only 15 metabolites displayed an absolute fold-change higher than 1 in at least one time-point. The untargeted metabolomics detected 11,190 ions but only a small percentage could be annotated to known metabolites (see section B.1.2). Briefly, the detected ions masses were mapped onto an existing library built using the iMM904 genome-scale metabolic reconstruction of yeast metabolism [Mo et al., 2009; Zomorrodi and Maranas, 2010]. Different natural modifications to the ions can occur and these have an impact in the detected mass and therefore their identification in the MS. To account for this a stringent filter was used to consider only deprotonated modifications. This allowed to rigorously annotate 196 ions which mapped to 270 yeast metabolites. As before, fold-changes were estimated compared to the initial time-point. Strong fold-changes were visible immediately after perturbation at time-point 25 seconds across the majority of the detected ions in salt and co-stimulation conditions only (Figure C.1 A). We hypothesised that due to the high concentration of salt added to the cell cultures, 0.4 M, MS measurements were affected because of an increased ionisation of the matrix. Considering this, to remove the possible matrix effect instead of normalising the experiments to time 0 seconds I normalised to time 25 seconds, since if there is a matrix effect it is already visible at 25 seconds. This comes with the drawback of discarding any fast changes happening between 0 and 25 seconds. Indeed normalising to time 25 seconds removed the strong and immediate changes (Figures C.1 A, 2.3 A, 2.3 B), therefore for the untargeted metabolomics I kept 25 seconds as initial condition and all fold-changes were computed compared to it. Consistently with targeted metabolomics few metabolites displayed.
Figure 2.2: Overview of the experimental design and metabolic and phosphorylation profiles of readouts of interest. A) Experimental setup of the three different stimulations, i.e. salt, pheromone and co-stimulation with both. Cells were inoculated in shake flasks and treated one hour before sample extraction with Cdc28 inhibitor. Intracellular metabolites were collected using fast filtration followed by cold-extraction. Targeted metabolomics was acquired with LC-MS/MS and untargeted with QTOF-MS. B) Agreement between LC-MS/MS and QTOF-MS and metabolites fold-changes. Metabolites and phosphosites profiles under salt, pheromone and co-stimulation. Ions mapping to more than one metabolite are marked with an asterisk (*).
strong changes, only 39 deprotonated ions had an absolute fold-change higher than 1 in at least one time-point. 11 unique metabolites were quantified with both methods and these showed a significant concordance (spearman’s rho=0.58, $p$-value=4.1e-31) (Figure 2.2 B).

Of the 39 measured ions, these include several examples of metabolites known to be regulated under these conditions (Figure 2.2 C). Considering the close proximity of fumarate and malate in the metabolic networks, it is reassuring that both showed strong similarity in salt and pheromone conditions. Glycerol 3-phosphate displays an accumulation overtime under salt stimulation consistently with known signalling regulation of Gpd1 leading to the production of glycerol [Saito and Posas, 2012]. Yeast cells also produce and accumulate trehalose under different types of stress conditions, including osmotic stress, and this is clearly visible with the trehalose profile [Hohmann, 2002; Saito and Posas, 2012]. While the metabolic implications of the pheromone stimulation in yeast are generally poorly understood the pheromone MAPK pathway is known to undergo regulation [Merlini et al., 2013]. TOR and MAPK signalling pathways have been shown to crosstalk [Brückner et al., 2011]. Therefore, it is interesting to see that metabolites involved in the biosynthesis of amino-acids, such as, L-glutamine, N-acetyl-L-glutamate and L-citrulline are significantly accumulated over time after pheromone stimulation. Some of these have been previously shown to directly influence Tor1 activity [Oliveira et al., 2015a].

### 2.2.2 Phosphoproteomics profiles of osmotic and pheromone response

The phosphoproteomics changes were measured using label-free MS after salt, pheromone and co-stimulation with both [Vaga et al., 2014]. Six time-points were measured, 0, 1, 5, 10, 20 and 45 minutes, where 0 minutes represent the unperturbed state right before adding the stimuli. All time-points fold-changes were estimated compared to the initial time-point and biological replicates were averaged. Multiple phosphorylated phosphopeptides were discarded and the remaining were mapped to the protein sequence to obtain information about the phosphosite absolute location. Only phosphopeptides measured at least twice in the three replicates were considered. Fold-changes were then estimated comparing the sample measurements to time-zero and significance was statistically
assessed using the biological replicates.

Regulatory sites, T174 and Y176, of Hog1 displayed increased phosphorylation and this is consistent with its expected increased activity (Figure 2.2 D) [Kanshin et al., 2015]. Pbs2 also known to interact with Hog1 [Martin et al., 2015] displays similar profile to Hog1 residues with increased phosphorylation. Sko1, known to be regulated by Hog1 downstream phosphorylation cascade, concurrently displays increased phosphorylation in both salt and co-stimulation conditions. Also the metabolic enzyme Gpd1 involved in the biosynthesis of glycerol [Saito and Posas, 2012] displays an early decreased phosphorylation in S24 (Figure 2.2 D) and this is concordant with the increased enzymatic activity expected upon osmotic stress. Moreover, Gpd1 S24 dephosphorylation is only transient and this has been discussed as a possible flexible mechanism in yeast to be able to respond to further osmotic imbalance [Mitchell et al., 2015].

To understand the global implication of phosphorylation in metabolism I calculated the number of residues that are significantly changing in phosphorylation across all metabolic enzymes (Figure 2.3). Under salt stimulation 8.08% (175 out of 2167) of the phosphosites were significantly differentially phosphorylated in at least one time point, out of these 13.14% (23) were mapping to metabolic enzymes. For pheromone stimulation, 12.96% (249 out of 1921) of the phosphosites were significantly changing in phosphorylation and out of these 10.84% (29) were residues from metabolic enzymes. For the co-stimulation 18.93% (367 out of 1939) of the phosphosites were significantly phosphorylated and 11.44% (42) were part of metabolic enzymes. Overall, both perturbations seem to alter different residues (Figure C.3 A, B), nevertheless this overlap is more pronounced when I consider the substrate protein (Figure C.3 C, D). Therefore, this may hint that both perturbations may crosstalk using different regulatory mechanisms [Vaga et al., 2014]. Salt stimulation displayed an early response with a majority of the measured phosphosites changing in phosphorylation within the first 10 minutes after perturbation, while the pheromone showed a homogenous change across the considered time frame with a slight increase towards the last time-points (Figure 2.2 A). Changes in the distributions of metabolic enzymes residues have a similar profile to all the rest and cover an average of 11.44% of the significantly changing phosphosites across the three conditions.

To understand the biological processes regulated I performed a pathway enrichment analysis with a hypergeometric test using KEGG pathways and the
Figure 2.3: Phosphorylation changes of the different conditions. A) Number of significantly changing phosphosites across the three different conditions. Phosphosites were considered to be changing significantly if they have in at least one condition an absolute fold-change higher than two and an adjusted \( p \)-value lower than 0.05. B) KEGG pathway enrichment analysis using a hypergeometric test. Proteins were considered if they had at least one significantly differentially phosphorylated phosphosite. C) Time profile of phosphorylation sites falling in proteins downstream of Tor1.
proteins with at least one significantly phosphorylated residue (Figure 2.3 B). The top enriched pathway under salt condition is termed *longevity regulation* and it involves Tor signalling pathway, such as Sch9, Msn2 and Msn4, these are important for nutrient regulation in yeast [Hohmann, 2002; Broach, 2012; Oliveira et al., 2015b]. Interestingly, these proteins displayed significant change only under salt stimulation conditions (Figure 2.3 C). Also, Tor and metabolism are tightly associated controlling the intake of amino-acids [Oliveira et al., 2015b]. *Starch and sucrose metabolism* pathway has significantly enriched, more predominantly in the pheromone condition, and this involves several metabolic enzymes involved in the usage of different sugars, such as, glucose, fructose and lactose (Figure 2.3 B). MAPK signalling pathway is also enriched in salt conditions, concordantly, with the increase phosphorylation and activity of Hog1 and downstream targets (Figure 2.3 B) [Chavel et al., 2014; Vaga et al., 2014]. This emphasised the extent of the phosphorylation adaptation undergoing under salt and pheromone perturbations and that metabolism is also regulated via several enzymes that show significant changes in phosphorylation.

### 2.2.3 Metabolic similarity between stimulation and co-stimulation

To understand the contribution of each stimulation to the phenotype measured in the co-stimulation I analysed the correlation of the significantly changing metabolites and phosphosites between each stimulation and the co-stimulation (Figure 2.4). For the phosphoproteomics data-set the existence of missing values may affect the correlation. Thus, only phosphosites measured across all the time points were considered, leaving 33.08% (215 out of 560) of the phosphosites changing significantly at least once in any of the three conditions. It was possible to see that specific groups of metabolites have a co-stimulation response similar to both salt and pheromone simulations and therefore these are metabolites possibly regulated in both conditions (Figure 2.4 A, B). In contrast, certain metabolites displayed a strong correlation to a specific stimulation, for instance, trehalose in the co-stimulation was strongly correlated with salt stimulation, highlighting the specificity of its accumulation to salt perturbation (Figure 2.4 A, B). These correlation patterns were also visible at the phosphoproteomics level, where groups of phosphosites seem to be specific of each condition and others have a similar
Figure 2.4: Analysis of the association between each stimulation and the co-stimulation. Associations were performed using a pearson correlation of each metabolite, ion or phosphosite across the different conditions and time-points. A) Targeted metabolomics. B) Untargeted metabolomics. C) Phosphoproteomics. 

response in both conditions (Figure 2.4 C). Phosphoproteomics measurements allowed to study the crosstalk between NaCl and pheromone stimulated signalling pathways [Vaga et al., 2014]. However, for the metabolomics the general absence of strong and robust changes and the reduced number of measurements limits the analysis to only a few metabolites.

2.2.4 Systematic identification of regulatory phosphosites in metabolic enzymes

As highlighted before phosphoproteomics measurements lack explicit functional information. To this end, I used the metabolomics experiments as a way to infer the activity of metabolic enzymes and associate with phosphorylation measurements to hypothesise functional regulatory phosphosites. Subsequently, I integrated the metabolomics and phosphoproteomics experiments using a genome-
Figure 2.5: Workflow to identify putative regulatory phosphorylation-sites in metabolic enzymes. A) Phosphosites in enzymes one or two reactions away from the metabolite were considered as features for the linear regression models. The metabolite fold-changes were used as observations and a linear model was used to estimate the coefficients $\beta$. An intercept and a noise term were used in each linear regression. B) Number of significantly phosphorylated sites mapping to each of the significantly changing metabolites.

scale metabolic reconstruction of yeast [Mo et al., 2009]. This allowed me to query the network for measured phosphosites in the neighbourhood of each metabolite (Figure 2.5 A). In other words, one can check the phosphorylation changes in the metabolic enzymes catalysing the metabolic reactions surrounding each measured metabolite. For each metabolite I built two linear regression models one for each metabolomics experiment. The observations of the linear models are the metabolite fold-change across the different conditions and time-points, and the independent variables, or features, are the measured phosphosites in the metabolic enzymes catalysing reactions one or two steps away from the metabolite (Figure 2.5 A). These linear regression models represent a simplistic linear representation that does not consider other regulatory events besides phosphorylation changes. Therefore, these models focus solely on the phosphorylation changes of the neighboring metabolic enzymes to explain the metabolite variation. Metabolites changes that can be significantly well explained with phosphosites changes alone are likely to indicate relevant functional associations between the two. Additionally, these models allow to account for combinatorial contribution of different residues to explain the metabolic changes, and therefore represent a flexible
framework to discover regulatory events only visible upon multiple phosphorylations, such is the case of Hog1 that requires double phosphorylation at T174 and Y176 to be fully active. Notwithstanding, linear regressions are limited to model additive effects while more synergistic effect would need to be captured with product terms.

For this analysis only metabolites and phosphosites that were changing significantly in at least one condition were considered. On average there are approximately three phosphosites mapping each of the significantly changing metabolites with the exception of some metabolites, such as trehalose and acetyl-CoA, that display high number of phosphosites in the neighbour enzymes. Thus to understand the impact the topology of the metabolic network in the identification of phosphosites in the vicinity of the metabolite, I correlated the number of measured phosphosites with the number of reactions associated with each metabolite. This represented a positive correlation but not significant (Figure C.4), thus suggesting that I am not biasing towards highly connected metabolites.

Figure 2.6: Results of the linear regression models. A) Linear regression adjusted coefficient of determination of each metabolite in each metabolomics data-set. B) Features coefficient for the top predicted metabolites. Values reported in the heatmaps are the coefficients trained by the linear regressions.

Considering the different number of phosphosites used to estimate the variation of each metabolite, I used and adjusted coefficient of determination (see section A.3.1) that measures the goodness of the fit and adjusts to the number of features (Figure 2.6 A). Glycerol 3-phosphate was among the top predicted metabolites concurrently with our previous expectation that glycerol synthesis is mainly regulated via phosphorylation changes. Furthermore, trehalose and tre-
halose 6-phosphate profiles were also significantly well explained by the phosphorylation changes of surrounding enzymes. The importance of each phosphosite to predict the metabolite fold-change was assessed using the coefficients estimated by the fitted linear model (see section A.3.1) (Figure 2.4 B). Among the most important features to explain glycerol 3-phosphate changes were Gpd1 S24, but the association is positive and therefore counter intuitive to what is expected, that is decreased phosphorylation renders the metabolic enzyme active and therefore increased concentration of its products. Moreover, upon closer look to the dynamic profile of glycerol 3-phosphate (Figure 2.2 C) and Gpd1 (Figure 2.2 D) it is possible to see that the accumulation of the metabolite occurs on later time points, while the dephosphorylation occurs almost immediately after the perturbation with salt. The reestablishing of the phosphorylation to allow the cells to react to another osmotic stress [Mitchell et al., 2015] hinders the association with the metabolite. Of note, S27 is also a known regulatory site of Gpd1 with similar regulatory pattern of S24, nevertheless it displays a weaker association. This is explained by the fact that the phosphosite does not have a decrease in phosphorylation as seen in S24 and also the measurements do not seem to be robustly measured across replicates. Another well predicted metabolite is trehalose for which several phosphosites falling in the trehalose synthase complex subunits Tsl1 and Tps3 [Reinders et al., 1997; Bell et al., 1998]. In particular, Tsl1 catalyses the synthesis of trehalose from glucose 6-phosphate, therefore it is important to see that Tsl1 S135 and Tsl1 T194 phosphorylation-sites display inverse associations between them. However, other residues display the same direction, such as, Tsl1 S79. While little functional information exists on the identified phosphosites [Sadowski et al., 2013] these residues are likely displaying a regulatory effect on the assembly and activity of the enzymatic complex.

2.3 Discussion

The advent of high throughput experiments has enabled the systematic and comprehensive measurement of large collections of molecular data [Bodenmiller et al., 2010; Schulz et al., 2014; Kemmeren et al., 2014]. The acquisition of different types of omics will require tailored and computationally efficient methods to integrate them and extract biological insights. Specifically, metabolomics and
phosphoproteomics have been integrated before to study regulatory phosphosites of metabolic enzymes in CCM [Oliveira et al., 2012].

In this chapter, I presented two novel metabolomics data-sets, targeted and untargeted, that characterise the intracellular metabolic changes in yeast upon salt, pheromone and co-stimulation with both. Metabolomics experiments were designed to pair existing phosphorylation measurements [Vaga et al., 2014]. As reported in the phosphoproteomics experiments, wild-type yeast cells displayed a long delay initiating the response to pheromone stimulation when compared to osmotic stress, which was immediate. To circumvent this problem and have both stimulations responding in similar time scales Cdc28 analog sensitive strains were used instead, and Cdc28 inhibition is performed one hour before stimulating the cell cultures. On the one hand, metabolomics experiments recapitulated previously reported responses for salt perturbation, i.e. accumulation of glycerol 3-phosphate and trehalose, and these were robustly measured in both data-sets. On the other hand, there was a general lack of strong and robust intracellular changes, thus leading to question the implication of Cdc28 inhibition. Cdc28 kinase is part of the cyclin-dependent kinase complex (CDK) and a main regulator of cell cycle [Mendenhall and Hodge, 1998], in particular Cdc28 has an important role in the initiation of the S phase and in the transition into M phase [Hartwell et al., 1974; Dirick et al., 1995; Bloom and Cross, 2007]. Considering that, inhibition of Cdc28 is expected to impact cell cycle progression and arrest cells in the G1 or G2 phase. While cells can be arrested in both stages it is intuitively to assume that the majority is arrested in the entry to the S phase due to the distribution of the cells along the different cell cycle stages. Therefore, one can not exclude that cell cycle arrest can have profound implications in the metabolic activity of the cells and consequently act as a confounding effect in the conditions analysed.

The untargeted metabolomics data-set had an added level of complexity due to the ionization of the salt stimulated samples where a matrix effect was visible in the MS runs. While this effect is taken into account by normalising to time 25 seconds all the changes that occur between 0 and 25 seconds were discarded as a consequence.

Pairing the phosphoproteomics data-set with metabolomics measurements across the same conditions allowed me to study the possible regulatory associations between phosphosites and metabolites. Several metabolic network recon-
structions are available for yeast [Mo et al., 2009; Zomorrodi and Maranas, 2010] and they cover metabolic reactions at a genome-scale level and provide information about enzymatic regulation of metabolism. This overlap between different biological processes is important and makes genome-scale models a natural resource to establish associations between modifications in metabolic enzymes, e.g. allostery regulation, post-translation modifications or regulation of expression, and changes in metabolism. Significant changes in metabolites concentrations are suggestive of changes in the activity of metabolic enzymes. Considering this, I used a genome-scale metabolic model [Mo et al., 2009] to define the metabolic enzymes phosphosites surrounding the changing metabolites within a radius of one and two reactions. Subsequently, I built linear regression models to estimate the metabolites changes across all the conditions based on the surrounding phosphosites of metabolic enzymes. These multilinear regression models allowed me to explore different combination of phosphosites that can be functionally related with the activity of the enzymes and thereby changing the metabolites concentration. Moreover, these associations can be performed systematically across all measured metabolites and phosphorylation-sites in a computationally efficient and mathematically robust framework.

As a response to the osmotic imbalance, yeast increases activity of Gpd1 via dephosphorylation of S24 and S27 and thereby increases the biosynthesis of glycerol (Figure 2.1). With the metabolomics measurements at hand it was not possible to measure robustly glycerol changes, only the intermediate glycerol 3-phosphate. Nevertheless, it would be expected that S24 was negatively associated with glycerol 3-phosphate. Contrarily to what was expected, the linear regression models showed that S24 is an important feature but it displays a positive association. The phosphorylation profile of S24 shows a strong dephosphorylation occurring in the first 5 minutes after salt stimulation, but its phosphorylation is re-established at 15 minutes. This behaviour has been explained has a reset mechanism of yeast to be able to respond to further osmotic stress [Mitchell et al., 2015] and can partially explain the incorrect association. Moreover, this raises the importance of the different time-scales of the biological processes that can hinder the integration of the different types of measurements. In general it was possible to estimate well the fold-changes of a limited number of metabolites. In the one hand, this reflects the lack of strong metabolic changes with a limited number of metabolites being considered for analysis. On the other hand, this
is also a consequence that only regulatory events by phosphorylation are being considered and therefore overlooking a whole set of other regulatory mechanisms, such as, post-transcriptional regulation. Considering that the experiments extend to a total of 45 minutes protein pool changes can not be discarded and therefore may affect specific proteins. This can be circumvented by focusing on earlier time-points or integrating proteomics measurements as a covariate for the linear regression models. Both options are not feasible on this particular data-set, since discarding later time-points reduces the already small number of samples and may hamper even further their predictive power, and considering the protein concentration change is not possible since no proteomics data-set under the same conditions was acquired.
CHAPTER 3

SYSTEMATIC ANALYSIS OF
TRANSCRIPTIONAL AND
POST-TRANSCRIPTIONAL
REGULATION OF METABOLISM IN
YEAST

In this chapter, I present a novel approach to infer regulatory protein-metabolites interactions in yeast. All the results in this chapter are the outcome of my work with critical input from David Ochoa and close supervision of Pedro Beltrão and Julio Saez-Rodriguez. Omar Wagih assembled the transcription-factor regulatory network. This work also benefited from fruitful discussions with Damien Arnol and Paolo Casale. The novel intracellular metabolomics data-set of yeast upon salt and pheromone was presented in Chapter 2 and was possible via an internship at ETH Zurich with help and close supervision of Zrinka Nakic, Mattia Zampieri and Uwe Sauer. The experimental data and the computational approach are the recipient of a manuscript submitted to a peer-reviewed journal (Gonçalves et al.).
3.1 Introduction

Cellular adaptations are achieved via combined responses from protein-protein to metabolic networks [Chubukov et al., 2014; Patil and Nielsen, 2005; Herrgård et al., 2006; Daran-Lapujade et al., 2007]. Despite their distinct molecular roles these biological networks interact with regulatory processes that allow them to coordinate their response. Several regulatory processes have been identified [Daran-Lapujade et al., 2007; Oliveira et al., 2015b] but the accurate and systematic assessment of their functional impact is still a challenge. Recent studies have explored regulatory interactions by using tailored computational approaches and combining different types of molecular measurements [Zelezniak et al., 2014; Gerosa et al., 2015]. A limitation to these studies is that, currently, large-scale experimental data-sets are still scarce [Kemmeren et al., 2014; Schulz et al., 2014; Bodenmiller et al., 2010].

High-throughput approaches provide comprehensive characterisation of the molecular changes but lack functional information. Mechanistic insights can be gained by performing pathway enrichment analysis and thereby providing a perspective of the biological processes or pathways involved in the cellular response. Nevertheless, it is generally the case that the exact definition of a pathway or biological process is ambiguous due to incomplete knowledge and high degree of interactions between the molecular species. Despite that, methods such as gene-set enrichment analysis (GSEA) [Subramanian et al., 2005] have proven to be useful for exploratory analysis of different types of molecular data. Thus, much efforts have been put in the manual curation of lists of proteins/genes characterising different biological processes, [Mootha et al., 2003; Subramanian et al., 2005; Kanehisa, 2004; Kanehisa et al., 2016], targets of transcription-factors (TFs) [Mathelier et al., 2014; Matys et al., 2006; Kemmeren et al., 2014; Hu et al., 2007; Chua et al., 2006] and targets of kinases and phosphatases [Hornbeck et al., 2015; Sadowski et al., 2013; Gnad et al., 2011]. Taking advantage of this manually curated information of target genes and phosphosites mathematical approaches have been developed to infer the enzymatic activity of TFs [Cheng et al., 2012; Schacht et al., 2014] and kinases [Casado et al., 2013; Mischnik et al., 2016; Ochoa et al., 2016]. These approaches provide valuable insight on protein activity status and can be used in integrative analysis to infer causal directed relationships.
itations of these methods are: (i) the associations between the regulatory protein and its targets may have different functional impact, i.e. activatory or inhibitory; (ii) biological responses are robust systems that often have compensatory effects, i.e. multiple regulators regulate the same target; and (iii) our knowledge of these associations is at best incomplete at the moment, thus one may wrongly infer the activity status of a regulator. Bearing these limitations in mind, one can still infer useful information about the activity of TFs and kinases/phosphatases [Casado et al., 2013; Cheng et al., 2012].

In the yeast *Saccharomyces cerevisiae* most of the integrative analysis of metabolism have focused on the role of transcriptional regulation [Patil and Nielsen, 2005; Zeleznjak et al., 2014; Gerosa et al., 2015; Oliveira et al., 2015a]. These studies overlook the fact that metabolism is also often regulated by PTMs, for example, phosphorylation is a key regulatory process to control protein enzymatic activity [Hohmann, 2002; Oliveira et al., 2012]. Analysis studying the regulatory relationships between signalling transduction and metabolism are much sparser [Yugi et al., 2014; Oliveira et al., 2012, 2015b].

In this chapter, I present a computational approach to systematically identify putative post-transcriptional and post-translational regulatory mechanisms of metabolism (Figure 3.1 A). I studied post-translational regulatory interactions under salt and pheromone stimuli by acquiring a time-resolved metabolomics data-set (see 2) and combining it with existing phosphorylation measurements [Vaga et al., 2014]. Furthermore, I integrated several experimental data-sets measuring transcription, phosphorylation and metabolite changes across 118 kinases/phosphatases knockouts [Kemmeren et al., 2014; Schulz et al., 2014; Bodenmiller et al., 2010] and under three different conditions perturbing nitrogen metabolism and Tor signalling pathways [Oliveira et al., 2015b] (Figure 3.1 A). In order to infer how transcription and phosphorylation changes are regulated I first estimated the *in vivo* enzymatic activity of TFs, kinases and phosphatases. To this end, I took advantage of prior-knowledge on regulatory interactions of gene expression and phosphorylation and of mathematical approaches that have been developed to infer the activity status of transcription factors [Cheng et al., 2012; Schacht et al., 2014] and kinases [Casado et al., 2013; Mischnik et al., 2016; Ochoa et al., 2016] (Figure 3.1 A). Enzymatic activities are hard to measure directly and provide functional information about the protein regulators involved in the cellular response. Regulators activities were integrated with the metabol-
Figure 3.1: Analysis and experimental design and data consistency. (A) Representation of the different types of data-sets used in the analysis. Transcriptomics and phosphoproteomics data-sets are used to estimate transcription factor and kinase/phosphatase activity changes, which are then separately associated with the respective metabolomics data-sets using multilinear regression models. (B) Experimental design used to acquire the intracellular metabolomics measurements. Cdc28 analog sensitive yeast strains inoculated in shake flasks were treated with the Cdc28 inhibitor. The unperturbed initial time points were taken 1 hour after the Cdc28 inhibitor and before adding the NaCl and pheromone. Sample filtration, metabolite extraction and MS injection were performed in parallel on the samples from independent biological experiments. (C) Metabolites fold-changes correlation between targeted and untargeted metabolomics. (D) Representative metabolite profiles of individual metabolomics experiments.
omics measurements using a machine learning approach to infer putative regulatory interactions. This approach could accurately estimate the activity status of known regulatory proteins, and identified protein-metabolite associations capable of robustly estimate the metabolic phenotype of previously unseen conditions.

3.2 Results

3.2.1 Compendium of yeast metabolic, transcriptional and phosphoproteomic paired data-sets

To study the interplay between metabolism and post-translational and transcriptional networks in yeast I compiled three multi-omics large-scale datasets containing genetic perturbations, i.e. kinases/phosphatases knockouts, and acute environmental perturbations for nitrogen related conditions, osmotic shock and pheromone stimulation (Figure 3.1 A).

I started by characterising the metabolic adaptations of yeast under osmotic and pheromone by acquiring dynamic intracellular metabolite measurements. To this end, I resorted to the metabolomics data-set presented in Chapter 2 that pairs a previously described phosphoproteomics data-set under these same conditions [Vaga et al., 2014]. For this chapter co-stimulation condition was discarded, thus only independent perturbation with salt and pheromone were considered. To summarise, the metabolomics measurements comprise time-solved changes across 45 minutes, 0 and 25 seconds and 1, 4, 5, 9, 10, 15, 20, 25, 35, and 45 minutes, where 0 seconds represents the time-point shortly before the stimuli are added. Cell material was extracted with fast-filtration and then analysed with targeted (LC-MS/MS) [Buescher et al., 2010] and untargeted (QTOF-MS) [Fuhrer et al., 2011] mass-spectrometry (see section B.1). In order to estimate the reliability of the metabolite measurements, I compared the metabolic fold-changes measured in both targeted and untargeted MS (Figure 3.1 C). A total of 11 unique metabolites were quantified with both methods and these showed strong concordance (spearman’s rho=0.77, p-value=1.9e-44) (see section B.1.2). On the untargeted data-set, 33 ions were defined as significantly changing in at least one of the time-points analyzed. The metabolomics measurements recapitulated previously reported biological observations and these were described in detail in
Chapter 2 section 2.2.1 (Figure 3.1 D).

The dynamic metabolomics and phosphoproteomics were then integrated with a compendium of different biological experiments to identify robust signalling regulatory interactions with metabolism across a variety of conditions. To this end, I considered a large panel of 115 kinases/phosphatases knockouts, for which molecular changes at the transcript [Kemmeren et al., 2014], phosphorylation [Bodenmiller et al., 2010] and at the metabolite [Schulz et al., 2014] level were characterised (Figure 3.1 A) (see section B.1.3). Dynamic measurements were also considered by integrating metabolomics, transcriptomics and phosphoproteomics data-sets measuring the response to three different perturbations around nitrogen metabolism and TOR signalling [Oliveira et al., 2015b,a]. In these studies, yeast cells were perturbed by varying the growth medium from poor to rich nitrogen growing conditions (nitrogen upshift) and vice-versa (nitrogen downshift). Yeast cells were also stimulated with Rapamycin, thereby inhibiting Tor1, a condition that resembles the nitrogen downshift (Figure 3.1 A). Assembling all the experimental data-sets together comprised a total of 143 different conditions for which metabolic, phosphorylation and gene expression measurements are available, except for salt and pheromone conditions where transcriptomics was not acquired.

3.2.2 **Inferring enzymatic activity of kinases, phosphatases and transcription-factors**

Changes in gene expression and in protein phosphorylation can be combined with metabolic measurements to identify possible regulatory associations. However, identification of functional regulatory interactions is hampered by the fact that expression is a poor proxy for TFs activity [Cheng et al., 2012; Schacht et al., 2014] and phosphorylation sites often display no functional impact in protein activity [Beltrao et al., 2012; Oliveira et al., 2012]. Therefore, to circumvent these limitations I estimated the enzymatic activity of TFs and kinases/phosphatases.

Enzymatic activity of kinases/phosphatases were estimated resorting to a comprehensive set of manually curated kinases/phosphatases-substrates interactions from PhosphoGrid [Sadowski et al., 2013]. TFs activities were inferred using a regulatory network obtained by combining gene-expression data from TF knockout experiments and TF binding sites from ChIP-chip experiments (see section A.2.2). The enzymatic activity of a regulator can be estimated by considering
the changes of its targets [Casado et al., 2013; Cheng et al., 2012]. For example, by analysing the phosphorylation changes of reported target sites of a protein kinase/phosphatase, one can predict whether the kinase/phosphatase is changing significantly (Figure 3.2 A).

GSEA [Subramanian et al., 2005] can be used to quantify and estimate the significance of the protein activity changes [Ochoa et al., 2016]. Nevertheless, the fact that it relies on a permutation scheme to calculate the activity score makes it a computationally demanding method with exponential complexity when compared to the number of measured phosphosites. Additionally, crosstalk, a characteristic phenomenon in signalling pathways, lead to kinases/phosphatases regulating overlapping phosphorylation-sites, thus increasing the similarity of the estimated activity score for proteins with similar sets of targets. Consequently, I propose a linear regression approach for estimating activities (Figure 3.2 A):

\[ Y = \beta X + \psi \] (3.1)

where, the dependent variable \( Y \) represents the phosphosites measurements across the sample; \( X \) is the connectivity matrix representing the associations between kinases/phosphatases extracted from the manually curated database, if \( X_{ij} \) equals to 1 it represents that phosphosite i is regulated by a kinase or phosphatase j, 0 otherwise; \( \psi \) represents the normally distributed error of the fit; \( \beta \) are the weights of the kinases/phosphatases and represent the activity scores. This linear formulation provides a computationally efficient framework that can be implemented in a variety of linear regression machine learning approaches, e.g. LASSO, Ridge or Elastic Net. Also it has the advantage to be a computationally inexpensive procedure with the capacity to scale-up to thousands of measurements, and partially accounts for compensatory effects between kinases/phosphatases that regulate overlapping phosphosites. This approach can be easily implemented in any programming language with support to already implemented machine learning approaches. In this thesis, I implemented it in Python using the machine learning module scikit-learn [Pedregosa et al., 2011]. I used the prior-knowledge regulatory networks to estimate the activity of 91 transcription factors and 103 kinases and phosphatases across all the conditions (see sections A.2.1 and A.2.2). The kinases activities scores estimated with the linear regression method were compared to an existing method [Ochoa et al., 2016].
activity profile of representative proteins for each experiment used in the analysis.

Figure 3.2: Protein activity analysis. (A) Representation of the workflow used to estimate protein activities using as input an experimental data-set and a regulatory network. Regulatory networks contain either the kinase/phosphatase-substrate interactions or TF-gene associations. A linear regression approach is used to estimate the activity of each TF and kinase/phosphatase in each condition, considering the regulatory network and the measured regulated genes or phosphosites, respectively. (B, C, D) Estimated activity profile of representative proteins for each experiment used in the analysis.
Protein activity profiles displayed strong agreement across all the samples (0.85 average Pearson correlation) (Figure C.5).

The phosphoproteomics data-sets contain 85.2%, 49.2% and 20.0% of missing values in genetic, nitrogen and salt/pheromone perturbations, respectively. For this reason, kinases/phosphatases activities could not be predicted in 3,227 (48.0%), 498 (25.2%) and 434 (7.7%) cases for the genetic, nitrogen metabolism and salt/pheromone perturbations, respectively. However the estimated changes in kinase activities do not rely on always measuring the same set of target phosphosites, hence the kinase activity changes are less sparse than the original datasets of phosphosite measurements. For the dynamic experiment of salt and pheromone stimulation there is no transcriptomics available thus transcription factors activity estimation was not possible.

Nitrogen downshift and rapamycin are similar conditions that inhibit Tor1 activity; in contrast, nitrogen upshift displays increased Tor1 activity. Thus, it is reassuring that the predicted protein activities tend to have similar changes in time for the nitrogen downshift and rapamycin condition, and opposite changes for the nitrogen upshift (Figure 3.2 B, C). Several of the predicted activities are in line with known condition dependent activity changes. Examples include the Tor mediated inhibition of Msn2, Msn4 and Gln3 TFs [Beck and Hall, 1999] (Figure 3.2 B) and the kinases Npr1 [Schmidt et al., 1998], Rim15 [Pedruzzi et al., 2003] and Yak1 [Martin et al., 2004] (Figure 3.2 C). Moreover, Hog1 and Pbs2, central kinases in the response to osmotic stress, display increased activity profiles [Vaga et al., 2014; Saito and Posas, 2012] (Figure 3.2 D). Similarly, the Ste7 MAPK kinase of the pheromone pathway is predicted to be activated during pheromone stimulation (Figure 3.2 D). These examples suggest that the TFs and kinases/phosphatases activities are well predicted and can be used to explore regulatory associations with metabolic changes.

### 3.2.3 Growth rate implications in intracellular changes

General effects in the cell, such as cell cycle and growth rate, can act as confounding factors when searching for regulatory associations between TFs and kinases/phosphatases and metabolic changes. In particular, gene expression changes, which upon different perturbations have been shown to be tightly correlated with growth rate due to changes in the distribution of cells over the cell
cycle phases [O’Duibhir et al., 2014; Brauer et al., 2008]. Considering that relative growth rate measurements are available for the genetic perturbations experiments and for each time point of the dynamic nitrogen metabolism experiments, I set out to assess how much of the variation in the data-sets can be explained by growth rate alone. To this end, I performed a Principal Component Analysis (PCA) analysis on TFs and kinases/phosphatases activities and metabolomics measurements. I then measured the correlation between relative growth rate and each of the top three principal components (PCs) (Figure C.6 A). Relative growth rate displayed a moderate correlation, pearson correlation coefficient of 0.25, with PC 1 (which explains 29.4% variation of the data) of the genetic perturbations metabolomics data-set. Growth rate displayed stronger correlations, 0.35 and -0.54, with PC 1 of kinases/phosphatases and TFs activities which explained 18.4% and 44.6% of the variation in the data, respectively (Figure C.6). The same analysis was performed for the dynamic nitrogen data-sets. Metabolomics PC 2 (17.1%) displayed a strong correlation with growth rate over time, pearson coefficient of 0.72. Kinases/phosphatases and TFs activities PC3 (10.9%) and PC2 (6.0%) showed also strong correlations with growth, pearson coefficient of 0.51 and 0.69, respectively (Figure 3.2). Of note, relative growth rate measurements seem to explain higher percentage of the variation in the steady-state genetic perturbations compared to the time-resolved nitrogen experiments.

For the subsequent association analyses I tested the impact of removing the growth rate from each data-set to rule out any confounding effects it may have on the identification of direct functional interactions. To this end, I regressed-out growth rates from the original metabolite measurements and estimated TFs and kinases/phosphatases activities using linear regression models and growth as a covariate.

3.2.4 Estimating metabolic changes from transcription-factors, kinases and phosphatases activities

Next I explored the correlations between TFs and kinases/phosphatases enzymatic activities and metabolic changes for each of the three experiments: genetic, nitrogen metabolism and salt/pheromone perturbations. To identify the relationships I used linear regression models that consider the estimated activities as features and metabolite fold-changes as observations. Considering the
low number of samples available, specifically for the time resolved experiments, a
cross-validation procedure of leave-one-out (LOO) was used (see section A.3.2).
This allowed me to understand how much information can be transferred within
each experiment to predict the metabolite variations in an independent testing
sample. Thus, for each experiment and each metabolite independent training
and test data-sets were generated leaving one sample out at a time for test, i.e.
single KO or time-point, and thereby generating a complete metabolomics matrix
with estimated fold-change values (Figure 3.3 A). The analysis is performed us-
ing TFs and kinases/phosphatases activities independently. In each experiment
four different types of input matrices are used to predict each metabolite, i.e.
kinases/phosphatases or TF activities with and without growth normalisation,
with the exception of the dynamic experiment with NaCl and pheromone for
which no growth rate and transcriptomics measurements were available. To min-
imize possible effects of over-fitting while training the linear models an Elastic
Net feature regularisation approach was used (see section A.3.2).

Firstly, I considered the genetic perturbations and assessed the capacity of TFs
and kinases/phosphatases activities to predict the changes of a given metabolite
across the panel of knockouts (Figure 3.3 B, metabolites). This would be, for
example, the changes in concentration of glutamine across the gene deletion con-
ditions. I evaluated the capacity of the models by correlating the independently
predicted fold-changes to the observed ones, this procedure was performed for
each metabolite and the results were summarized as correlation distributions.
The metabolites changes across the different conditions were generally poorly
predicted using either the TFs or kinases/phosphatases activities, displaying me-
dian correlations close to zero (Figure 3.3 B). This did not change when I used the
growth corrected data. I then measured how well the models predict the changes
in all metabolites in a given condition (Figure 3.3 B, conditions). This tests the
capacity to, for example, predict the changes of all metabolites in Hog1 knockout.
Overall, similar results as in the metabolites analysis were obtained, one differ-
ence is the improvement in predictive power considering kinases/phosphatases
activities. However, this increase is mostly lost when I use the growth corrected
data.

A similar analysis as with genetic perturbations was applied to the dynamic
experiments to estimate the metabolic variation. The trained models displayed
in general higher predictive power than the genetic perturbations (Figure 3.3 C).
The measured values are distributions of the correlation values between predicted and measured metabolite fold-changes. Delta- change. Independently predicted metabolite fold-change matrices were then correlated metabolite and condition wise with metabolomics. Changes in activity for Kinases/Phosphatases and transcription factors were used independently to estimate the metabolite fold-change. For each metabolite measured a multilinear regression analysis was performed using leave-one-out cross-validation of the analysis. (A) Diagram of the analysis. For each metabolite measured a multilinear regression analysis was performed using leave-one-out cross-validation. (B, C) Distributions of the correlation values between predicted and measured metabolite fold-changes with TFs and Kinases/Phosphatases activities. (A)

\[ \text{Predicted} = \varepsilon + \beta. \]
Overall, in the dynamic nitrogen experiments, TFs displayed better agreement between measured and predicted metabolite fold-changes than kinases/phosphatases, across metabolites (Figure 3.3 C, metabolites) and across conditions (Figure 3.3 C, conditions). Also, models trained with growth normalised activities obtained similar results to non-normalised data-sets. The metabolic changes in the salt and pheromone experiment could be reasonably explained using the kinases/phosphatases activities across metabolites (Pearson’s r 0.44) (Figure 3.3 C, metabolites) and conditions (Pearson’s r 0.39) (Figure 3.3 C, metabolites). These were generally worse than the nitrogen experiment, and this could be a consequence of the lower number of significantly changing metabolites when compared to the nitrogen experiments.

The predictive difference between growth normalised and non-normalised kinases/phosphatases activities in the genetic perturbations (Figure 3.3 B, conditions) seem to suggest that associations important to predict a new condition are generally dependent on global growth effects, and thereby likely to be indirect. Furthermore, the different predictive power between TFs and kinases/phosphatases on the dynamic nitrogen experiments seem to suggest that changes in TF activities are more predictive of metabolic changes. Nevertheless, one needs to consider that very different technologies are used to measure the underlying data-sets, i.e. transcriptomics and phosphoproteomics. and this may impact the predictive power of the data-sets.

3.2.5 Inferring putative regulatory protein-metabolites interactions

Considering that the metabolic predictions based on time-resolved experiments partially circumvented indirect effects and displayed the best predictive power (Figure 3.3 B, C) I decided to focus on these data-sets for the downstream analysis. Moreover, since growth has been shown to possibly act as a confounding effect (Figure 3.3 B) I only used the data-set with growth normalised for the nitrogen metabolism experiments. I also considered TFs and kinases/phosphatases separately and searched for putative regulatory associations with the metabolite changes.

I started by investigating the capacity of the nitrogen TFs activities to estimate the metabolites fold-changes in each perturbation. To this end, we used a
Figure 3.4: Overview of the putative protein-metabolite regulatory interactions. (A) Distribution of the metabolite predicted and measured correlation coefficients using a 3-fold cross-validation leaving each condition out at a time. (B) Heatmap of the TFs-metabolites associations where values represent the averaged coefficients. Asterisks identify significant correlations, FDR < 5%. (C) Correlation distributions between predicted and measured metabolite fold-change across all conditions. (D) Heatmap of the kinases/phosphatases-metabolites associations where values represent the varied coefficients. Coefficients are calculated using a bootstrap cross-validation randomly leaving 20% of all the samples out. This procedure is performed twenty times and averaged. Coefficients distributions are calculated using a 5-fold cross-validation leaving each condition out at a time. (D) Heatmap of the kinases/phosphatases-metabolites associations where values represent the varied coefficients. Coefficients are calculated using a bootstrap cross-validation randomly leaving 20% of all the samples out. This procedure is performed twenty times and averaged. Asterisks identify significant, FDR < 5%, Pearson correlations between the activity profiles and the metabolite fold-change across all conditions.
learning procedure, analogous to the one used before, but instead of leave-one-out, a three-fold cross-validation was used to leave each of the environmental perturbations out at a time (see section A.3.3). This was performed independently for each metabolite and the agreement between the measured and predicted values was calculated using Pearson correlation coefficients (Figure 3.4 A). Consistently with the previous analysis, a large fraction of the metabolites were well predicted in downshift and rapamycin conditions. The best performances are obtained in the nitrogen downshift and the rapamycin experiments, both displaying similar median correlations. This could be expected since these are related conditions and the relationships learned from one may be more readily applied to the other. Then, I considered only the best predicted metabolites (Figure C.7 A) and explored putative protein-metabolite associations using all the three nitrogen conditions together with bootstrapped linear regression models (see section A.3.3). The associations were estimated 20 times with 80% of the samples randomly selected, therefore generating 20 coefficients for each TF-metabolite association. The average of the TF-metabolite coefficients represents a confidence score on the association (Figure 3.4 B).

From the reported associations, Leu3 involved in the biosynthesis of leucine is positively associated with several metabolites involved in the biosynthesis of amino-acids, e.g. L-glutamine and L-citrulline [Friden and Schimmel, 1988; Nielsen et al., 2001]. Also, the involvement of Put3 in the proline utilisation pathways and its positive association with L-proline is captured by the linear models coefficients [Huang and Brandriss, 2000; Axelrod et al., 1991; Siddiqui and Brandriss, 1989]. These results seem to confirm that the regulatory interactions found are biologically relevant, although they can be a result of direct or indirect associations. For example, a direct interaction can occur if a TF regulates the expression of metabolic enzymes and thereby controls directly metabolite concentration. The association can also occur in the opposite direction where metabolites can directly regulate the activity of TFs. In contrast, indirect associations can be established, for instance, if metabolite changes are a consequence of downstream effects of TFs or if a cell state results in changes of both TF activity and metabolite concentration independently. In order to study this, I firstly identified the enzymes that use or produce each measured metabolite and considered a list of known TF-target proteins (from our assembled TF regulatory network), TF-gene genetic interactions (from BioGRID [Chatr-Aryamontri et al., 2015]) or
TF-gene functional interactions (from STRING [Jensen et al., 2009]) (see section A.3.3). I then searched for enrichment of known TF-target, TF-gene genetic and functional associations among the top predicted TF-enzyme-metabolite interactions (Figure C.8 A). No significant association was found. I also note that the variation in TFs activities are almost fully explained by the first PC that captures 85.6% of the total variance in the data (Figure C.5). Furthermore, TFs activities showed similar profiles within Tor1 inhibition conditions, nitrogen downshift and rapamycin, and opposing profiles in Tor1 activation condition, nitrogen upshift. Hence, this shows lack of specificity in the gene expression response and can partially explain the limited capacity to identify direct associations with metabolites. However, regressing-out the first principal component from the TFs activity scores and from the metabolomics measurements did not improve the enrichment in direct TF-target associations. These findings support the idea that although the TFs activities are predictive of metabolic changes these relationships are likely to be indirectly due to changes in cellular states or via transcriptional regulation of genes that are not those immediately in the vicinity of the associated metabolites.

For the kinases/phosphatases-metabolite associations I used all five dynamic perturbations: nitrogen upshift, nitrogen downshift, rapamycin, NaCl and pheromone (Figure 3.4 C). With the exception of the pheromone the other conditions showed similar median correlations between the measured and predicted metabolites fold-changes. However, the performance is overall lower than for the leave-one-out test (Figure 3.3 C), as would be expected from a more stringent evaluation. The top predicted metabolites were selected (Figure C.8 B) and an analogous approach used for the TFs was used to identify kinases/phosphatases-metabolite associations (see section A.3.3) (Figure 3.4 D).

Rim15 and Tpk1 displayed strong associations with the metabolites and these play a key role in the regulation of the cellular growth and their adaptation to nutrient availability [Chavel et al., 2014; Conrad et al., 2014; Broach, 2012]. Tpk1 inhibits the activity of Rim15 to regulate cell cycle, thus this justifies that both display opposite associations (Pearson’s r -0.61, p-value 2.7e-2). Furthermore, Rim15 is inhibited by Tor1 [Swinnen et al., 2006; Broach, 2012] and considering that L-proline is a poor nitrogen source leading to decreased Tor1 activity, this is consistent with the positive association between Rim15 and L-proline, and that Tpk1 displays the inverse. Of note, Tor1 and Rim15 display similar metabolite relationships despite their inverse biological association, this happens because
Tor1 activity is wrongly estimated due to lack of robustly measured targets. This is emphasised with the non-significant negative correlation between Tor1 and Rim15 activity scores (Pearson’s $r = -0.35$, $p$-value $6.86e-2$). The associations may be direct causal kinases/phosphatases-enzyme-metabolite relationships, but they could also be indirect or in the opposite direction where a metabolite change impact kinase activity. I performed an enrichment analysis similar to the one described before, but now considering from BioGRID genetic and physical interactions. For each kinases/phosphatases-gene network I tested for enrichment of true interactions in the top-predicted kinase-enzyme-metabolite associations (Figure C.8 B). A weak enrichment for functional interactions was visible (AROC=0.55, Figure C.8 B) and weaker still for direct kinases/phosphatases-target relationships (AROC=0.54, Figure C.8 B). These results suggest that the retrieved associations contain some direct kinases/phosphatases-target relationships but many are likely to be indirect.

3.3 Discussion

Signal transduction is an important cellular mechanism that allows cells to sense and respond to environmental cues. These mediate intracellular adaptations by regulating a variety of biological processes, including, metabolism and gene expression. Thereby interactions among different biological processes occurs and are very important to coordinate the whole phenotype of the cell. Nevertheless, the systematic identification and functional annotation of these regulatory interactions is still a challenge. Experimental data-sets covering different omics in similar conditions are becoming increasingly available and it is likely that analyses like the one I presented here will be useful to systematically explore these regulatory events.

The key novelty of the approach proposed here is that regulatory interactions are inferred from estimated enzymatic activities of TFs and kinases/phosphatases, which are difficult to measure directly. This provides the possibility of considering the activity profile of regulatory proteins, which has not been considered in previous studies of metabolism using phosphoproteomics and transcriptomics.

The results presented in this chapter suggest that it is possible to use kinases/phosphatases and TFs activities to predict changes of several metabolites...
in time-resolved experiments. However, the predictive power does not extend to all conditions. For example, the models trained with kinases/phosphatases activities showed limited capacity in the pheromone perturbation experiment. This can arise from the higher technical variability in the data obtained and also be due to lower number of regulated metabolites. Additionally, the regulatory interactions are often condition specific. As such, if the proteins that are important to regulate the nitrogen or osmotic related conditions are not used for the pheromone response, then the associations learned cannot be predictive of the pheromone induced metabolic changes. Interestingly, protein-metabolite interactions inferred from the genetic perturbations experiment displayed poor predictive power to estimate the metabolic changes of a new condition, in contrast to the dynamic experiments (Figure 3.3 B, C). This suggests that time-resolved experiments provide a more efficient design to infer regulatory associations by circumventing general confounding effects that may be present in the steady-state.

Nevertheless, while the inferred protein-metabolite interactions provide reasonable power to predict metabolic changes (Figure 3.4 A, C) of unseen conditions the predicted regulator-enzyme-metabolite interactions are not strongly enriched in previously regulatory interactions. Some features, particularly kinases or phosphatases activities, such as Rim15 and Tpk1, were important to estimate metabolites fold-change. This reassuringly assesses that the estimated protein activities profiles are biologically relevant and useful for inferring metabolic adaptation in novel conditions. The time-resolved metabolomics experiment under salt and pheromone resulted in only moderate metabolic changes, when compared to the nitrogen conditions. This smaller variation may explain the lower power in identifying regulator-metabolite associations in these conditions. This emphasises the importance of designing experiments that adequately perturb both signalling and metabolism, without possible confounding effects, such as Cdc28 inhibition. Another possible limitation of this approach is that, while I used comprehensive resources, I only considered prior knowledge of reported kinases/phosphatases-substrate and TFs-gene regulatory interactions. Furthermore, the lack of missing values in transcriptomics data-sets provides increased robustness to TFs activities when compared to kinases/phosphatases activities. Nevertheless, both protein activity profiles showed comparable predictive power, thus partially guaranteeing that the existing bias does not penalise greatly the kinases/phosphatases. The increasing number of recorded interactions for reg-
ulators will provide important information to expand the coverage of TFs and kinases/phosphatases for which it is possible to estimate activities and increase the robustness of the estimation.
CHAPTER 4

SIGNALLING REGULATION OF METABOLISM IN HEREDITARY AND LEIOMYOMATOSIS RENAL CELL CANCER

In this chapter, I introduce a novel computational framework to identify regulatory interactions between protein kinases/phosphatases and metabolic enzymes. All the results in this chapter are the outcome of my ideas and work with close supervision of Julio Saez-Rodriguez. The experimental data-sets presented were the result of a close collaboration with Marco Sciacovelli, Sofia Costa, Isaac Johnson, Pedro Cutillas and Christian Frezza. This work benefited from fruitful discussions with Pedro Beltrão. The experimental data and the computational approach are the subject of a manuscript in preparation (Gonçalves et al.).

4.1 Introduction

Cancer cells in general arise by an abnormal accumulation of somatic mutations in the genome that drive a complex and profound alterations of the cellular phenotype [Stratton et al., 2009]. Current technological advances made possible to molecularly characterise hundreds of tumours and cancer cell lines at different levels [Hakimi et al., 2016; Garnett et al., 2012; Iorio et al., 2016]. Specifically,
genomics allowed to understand the mutation landscape associated with particular types of tumours and the mutations occurring during tumour development [Stratton et al., 2009; Forbes et al., 2015; Martincorenna and Campbell, 2015]. While large-scale genomics and transcriptomics studies are becoming available for a wide variety of tumour types, (phospho)-proteomics and metabolomics are still lagging behind [Zhang et al., 2014; Moghaddas Gholami et al., 2013; Lawrence et al., 2015; Hakimi et al., 2016].

Figure 4.1: Diagram depicting the different biological measurements acquired and used to model signalling and metabolism in HLRCC cell lines. Signalling events are represented by changes in phosphorylation of phosphosites and integrated into a comprehensive kinase/phosphatase-substrate network. Metabolism is modelled using a genome-scale reconstruction constrained with consumption/release rates of metabolites into the medium (blue arrows) and intracellular fluxes are estimated using a sampling approach (dark and light blue distributions). Light blue represents HLRCC cell lines with FH mutated and dark blue represents the same cell line with FH expression reconstituted. The dashed line between FH mutation and the phosphorylation changes depict the indirect association between them.

Hereditary Leiomyomatosis and Renal Cell Cancer (HLRCC) tumours display a characteristic mutation of the metabolic enzyme Fumarate Hydratase (FH). FH catalyses the conversion of fumarate to malate, a reaction that takes place in the TCA cycle. This mutation leads to the impairment of the catalytic activity of
the enzyme and thereby to the accumulation of its substrate, fumarate. Recent work showed that FH deficient kidney cancers have decreased levels of AMP-activated protein kinase (AMPK) [Tong et al., 2011], alterations in HIF pathway [Sudarshan et al., 2007; Frezza et al., 2011], and that HLRCC derived cell lines morphologically display peripheral lysosomes a known marker for MTOR activity [Efeyan et al., 2012]. This suggests that signalling, as well as metabolism, undergo adaptation in HLRCC, although the mechanisms underpinning these adaptations are still poorly understood.

In this chapter, we characterised signalling and metabolic changes in HLRCC cell lines by designing and acquiring phosphoproteomics, proteomics and metabolomics measurements. These data-sets allowed me to study regulatory associations between signalling and metabolism using a novel computational framework to integrate phosphoproteomics with in silico estimated metabolic flux rates. This framework systematically identified putative regulatory interactions between protein kinases/phosphatases and metabolic enzymes (Figure 4.1), and specifically hypothesised a possible regulatory mechanism between ABL1 tyrosine kinase and PDHA1 metabolic enzyme.

4.2 Results

4.2.1 Molecular characterisation of HLRCC cell lines

HLRCC tumours are characterised by a hereditary mutation in the mitochondrial enzyme FH, rendering the enzyme inactive and leading to the accumulation of its substrate, fumarate. Abnormal concentrations of fumarate have been associated with alterations in signalling pathways [Tong et al., 2011; Efeyan et al., 2012; Frezza et al., 2011]. To study these molecular alterations we used cell lines derived from a HLRCC tumour, UOK 262. The tumour cell line phenotype was compared to the same cell line with FH reconstituted, UOK262pFH [Frezza et al., 2011] (see section B.2.1). These models allow to study the molecular adaptations underpinning the loss-of-function of FH in HLRCC. For brevity throughout the chapter UOK 262 cell lines will be addressed as KO and UOK 262 pFH as WT.

We first analysed the implication of FH mutation in the metabolic requirements of the cells by quantifying the consumption and release rates of metabolites to the medium (Figure 4.2 A) (see section B.2.1.2). Briefly, extracellular
Figure 4.2: Experimental data acquired for UOK262 (KO) and UOK262 pFH (WT) cell lines. A) Quantitative rates of consumption and release of metabolites to the medium. B) Phosphopeptide differential phosphorylation analysis between KO and WT. C) Proteomics differential abundance changes between KO and WT.

metabolites were measured with LC-MS where cells were plated and 24 hours later metabolites levels were acquired and compared to baseline levels measured in fresh media. The time difference and an approximate cell count were also taken into account to quantify the consumption/release (CORE) rates in mmol/gDW/h (see section B.2.1.2). Quantitative assessment of the rates is important as it discards possible general effects of cell size and provides measurements adequate to model metabolism with genome-scale models. CORE measurements have been acquired for KO and WT cell lines in two independent biological replicates with five technical replicates each. The experiment allowed us to fully quantify the consumption/release rates of 22 metabolites. In general there was a lack of significant differences between KO and WT (Figure C.9). An unsupervised hierarchical clustering showed that it was not possible to cluster KO and WT separately (Figure C.10), this is likely due to the general lack of changes in the metabolic rates. The different biological and technical replicates were averaged and, concordantly to the Warburg effect, KO cells displayed increased lactate secretion (Figure 4.2 A). A non-significant increase of glucose consumption is seen in KO cell lines (Figure C.9). Interestingly, glutamine consumption is observed in both cells with a higher intake in KO cell lines. This possibly suggests that glutamine is used
as an alternative carbon source to pyruvate for the TCA cycle, a phenomenon present in other tumours [Wise and Thompson, 2010].

We then set to measure the proteome and phosphoproteome of KO and WT cell lines with label free MS (see section B.2.1.3). Three biological replicates were acquired for the proteomics and for the phosphoproteomics three biological replicates with two technical replicates each were acquired. Subsequently, I performed differential analysis between the two conditions and estimated the statistical significance by taking the replicates variance in consideration using Limma R package [Ritchie et al., 2015]. Protein and phosphopeptide fold-changes were calculated using the mean difference of the replicates between KO and WT conditions for the proteomics and phosphoproteomics data-sets, respectively (Figure 4.2 B, C).

The measured proteome covered a total of 1,523 unique proteins and 11.88% (181) showed a significant change in concentration (FDR <5%). In agreement with the FH knockout, FH was underexpressed in KO cell lines, and VIM was identified as a top expressed protein in the tumour cell lines (Figure 4.2 B). VIM is an EMT driver gene and this corroborates previous results showing increased expression of EMT related genes [Frezza et al., 2011]. Proteomics showed a reasonable agreement with the RNA-seq transcriptomics measurements available for the same cell lines (spearman’s rho (r) of 0.55, p-value=9.7e-109) (Figure C.11). As expected, some proteins displayed a disagreement between the protein abundance and the transcript expression, these likely reflect different types of regulatory mechanisms occurring at post-transcriptional and post-translational levels. Reproducibility of the measurements was assessed with unsupervised hierarchical clustering where replicates showed higher correlation coefficients than all the pairwise comparisons (0.93 vs 0.87) (Figure C.12).

The phosphoproteome analysis covered 2,022 unique single phosphorylated phosphopeptides, mapping to 942 unique proteins, and 37.04% (749) of the phosphosites (mapping to 457 unique proteins) were significantly changing in phosphorylation (FDR <5%) (Figure 4.2 C). An unsupervised hierarchical clustering of the samples pearson correlation showed that the replicates of KO and WT cell lines group separately (Figure C.13). Similarly to the proteomics measurements, VIM also showed a significant increase in phosphorylation in the KO cell lines, although these changes are likely driven by the increase in protein abundance. The amount of significant changes seem to be greater in the phosphoproteome when
compared to the proteome (Figure 4.2 B, C). Nevertheless, one needs to consider the different number of replicates between the two data-sets, which make the comparison unfair. The phosphoproteomics data-set has twice more replicates and thereby is more likely to produce stronger p-values. Thus, considering the average fold-changes 5.44% (110) of the phosphopeptides and 2.89% (44) of the proteins have an absolute value higher than one. This also suggests greater changes in phosphorylation compared to the proteome. Still, protein intensities are estimated by averaging all the peptides that unambiguously map to a protein, thus one can not exclude that this will decrease the impact of outliers and ultimately provide smaller effect sizes. The same effect is not visible in the phosphoproteomics data-set since it is processed at the phosphosite level, for which generally there is only one measurement. Some metabolic enzymes displayed significant changes between KO and WT, in particular PDHA1 and GAPDH (Figure 4.2 C). Specifically, 54.76% (23/42) of the phosphopeptides in metabolic enzymes display significant changes (FDR $< 5\%$), and these map to 17 unique enzymes, out of the 30 covered. This supports the idea that metabolism is actively regulated by phosphorylation.

### 4.2.2 Genome-scale modelling of metabolism

Despite the fact that CORE experiments displayed a limited variation between the two cell lines, the significant increase in lactate secretion in KO cells emphasise that to some extend the metabolic network is regulated by FH mutation. Therefore, to characterise these changes in intracellular metabolic pathways I estimated the flux distributions of metabolic reactions by constraining a human genome-scale model [Duarte et al., 2007] with the CORE measurements.

Genome-scale models can be used together with Flux-Balance Analysis (FBA) [Orth et al., 2000; Varma and Palsson, 1994] to estimate the steady-state flux rates of the metabolic reactions. FBA based approaches require in general an optimisation function that represents the molecular objective of the cell. To this end, maximisation of the biomass has been shown to achieve accurate estimations of the cellular metabolic state [Schuetz et al., 2012; Varma and Palsson, 1994]. Nevertheless, for human cells this assumption overlooks many of the intrinsic molecular characteristics of the cells, which display specific roles that are often not easily representable. To circumvent the lack of a precise objective function I
resorted to a sampling approach based on Markov Chain Monte Carlo (MCMC) methods. This approach does not require an objective function, since instead of providing a single estimation of the metabolic flux rates that maximises a certain objective function, it provides samples of the possible flux distributions of the reactions based on the model at hand. Thus, it allows me to perform an uniform sampling of the flux distributions of the intracellular metabolic reactions considering the CORE measurements of each cell line. To ensure an uniform sampling of the solution space this approach requires thousands of iterations and therefore is computationally demanding. Considering that, I used an efficient and optimised implementation of MCMC methods for genome-scale metabolic models termed OptGpSampler [Megchelenbrink et al., 2014] (see section A.4).

A general genome-scale metabolic model [Duarte et al., 2007] is used to generate specific constrained models for KO and WT cell lines separately. FH loss-of-function mutation in KO cells was represented by limiting the flux rate of its catalysed reactions to zero, while in the WT cells they remained unaltered. The composition of the cell medium was also used to restrict the metabolites available for consumption. Despite the fact that the medium is not completely characterised due to the presence of bovine serum, this simplification ensures that differences in the sampling between KO and WT arise from metabolites that were previously identified. The metabolic models were then constrained using the CORE rates, hence generating two context-specific metabolic models (see section A.4) (Figure 4.3 A). Consistently with the measured metabolic rates, the metabolic models show small differences in general between KO and WT, with the exception to the secretion of lactate. The metabolic models also display high secretion rates of hydrogen, which are likely associated with lactate. The constrained models can then be used to sample the flux distributions of the internal metabolic reactions, thereby generating two flux distributions per reaction, i.e. KO and WT. This allows me to explore the intracellular metabolic differences between the two cell lines, which are estimated solely on the constraints defined before.

To partially assess the capacity of the metabolic reconstructions to estimate the rates of the exchange reactions I performed a leave-one-out cross-validation test. To this end, for each condition one metabolite was left out at a time and the rest were used to constrain the metabolic model. Then the estimated rate for the metabolite used for test was compared to the actual measurement. A good
agreement was seen between the predicted and the measured rates (spearman’s rho 0.65, $p$-value of 6.2e-03) (Figure 4.3 B). The models were able to estimate reasonably well the glucose intake rates, but were incapable of predicting the secretion of lactate in both KO and WT, predicting no secretion or consumption of lactate. This goes in line with the Warburg paradox which is not possible to explain solely from the metabolic point-of-view [Warburg, 1956; Vander Heiden et al., 2009; Basan et al., 2015].

### 4.2.3 Systematic analysis of associations between phosphorylation and flux changes

The differential phosphorylation analysis provided information of the phosphosites in metabolic enzymes that change significantly and are likely to be regulated. However, information about the functional implication of the residues is generally scarce and specifically for metabolic enzymes only a very small number have been annotated [Hornbeck et al., 2015]. To systematically analyse the activity status of the metabolic enzymes I used the estimated flux distributions of their catalysed reactions and assessed for each reaction the significance of the flux dif-
ferences between the KO and WT. Thereby, flux changes can be used as a proxy for the functional activity status of the metabolic enzymes and consequently associated with phosphorylation changes of the enzymes residues to find putative regulatory phosphorylation-sites. To this end, I considered all the reactions for which there is at least one phosphopeptide measured on the enzymes catalysing the reaction. Then, I systematically estimated the statistical significance of the mean differences between KO and WT for the flux distributions. A nonparametric alternative to the paired t-test, Wilcoxon test, was used since reactions fluxes distributions may follow a non-gaussian distribution. Wilcoxon p-values were then adjusted with Benjamini-Hochberg false discovery rate.

A total of 30 unique and significantly changing (FDR <5%) associations between reactions and phosphopeptides were obtained (Figure 4.4 A). These associations cover a total of 21 unique reactions, 16 unique enzymes and 18 unique phosphorylation-sites. The top changing reaction is a sodium transport reaction R_NAt5 followed with NAt3_1 both catalysed by SLC9A1 enzyme (Figure 4.4 B). This finding suggests a plausible mechanism to balance the increased hydrogen levels due to aerobic glycolysis using this sodium/ammonium proton transport to the medium. The metabolic fluxes are associated with the phosphorylation changes of SLC9A1 in the serine residues in position 693 and 788. Therefore, this indicates a possible signalling regulatory mechanism of metabolism transport reactions. The systematic analysis of the reactions that display significant flux changes between KO and WT comprises a list of putative regulatory points in the metabolic network that undergo regulation upon FH mutation (Figure 4.4 B).

4.2.4 Identification of kinases/phosphatases involved in the regulation of metabolic enzymes

The analysis presented before can be used to identify putative regulatory phosphosites in metabolic enzymes. Nonetheless, it is also important to understand which kinases and phosphatases are required to mediate these phosphorylation changes. Differential phosphorylation analysis provide a snapshot of the phosphoproteome changes and thereby identify processes that undergo signalling regulation. Nevertheless, this type information provides no direct insight into the activity status of the regulating signalling proteins unless integrated with prior knowledge. To
Figure 4.4: Associations between phosphosites in metabolic enzymes and their catalysed reactions. A) Enzymes phosphopeptides fold-changes mapped against the intracellular flux changes of the catalysed reactions. Associations listed in the legend are sorted by the significance of the flux change, where dark blue represents stronger changes. B) Distributions of all the reactions for which the catalysing enzymes have at least one phosphosite changing significantly. Violin plots were cut to fit the range of the observed flux rates.
this end, I assembled the regulatory targets of kinases/phosphatases from Omni-
Path [Türei et al., 2016], a comprehensive resource of kinases and phosphatases
substrate interactions. To assemble the kinase/phosphatase-substrate network I
considered: PhosphositePlus [Hornbeck et al., 2015], PhosphoELM [Dinkel et al.,
2011], HPRD [Keshava Prasad et al., 2009], Signor [Perfetto et al., 2016] and
DEPOD [Duan et al., 2015] resources (see section A.2.3).

Figure 4.5: Kinases and phosphatases activities. A) Correlation of ki-
nases/phosphatases activities between linear regression and GSEA approaches. B)
Kinases and phosphatases with an absolute GSEA activity score higher than 0.5, plot-
ting 51 out of 113.

The network contains 8,533 unique phosphorylation-sites, mapping to 2,364
proteins, and 548 kinases/phosphatases, with at least one reported substrate.
Kinases/phosphatases have a median of 5 substrates, and each substrate has a
median of 1 regulatory protein (Figure C.14 A, B). Only 2.8% (239) of the network
substrates were measured in the phosphoproteomics data-set, thereby allowing to
estimate the activity of only 24.3% (113) of the reported kinases/phosphatases.
The activities were estimated using the linear regression approach that I presented
in Chapter 3 and were also compared to the GSEA approach [Ochoa et al., 2016].
Both displayed very similar results (pearson’s rho 0.69, p-value 2.6e-17) (Figure
4.5 A). Considering that GSEA, in a different context, has been experimentally
validated to estimate the activity of Aurora kinase A (AURKA) [Ochoa et al., 2016] this approach was used for the downstream analysis. The top changing kinases/phosphatases (Figure 4.5 B) show the increased activity of ROCK1 in KO cells. Interestingly, this is a protein associated with cytoskeleton and cell motility and thereby possibly relevant to the metastatic phenotype of KO cells [Maekawa et al., 1999; Hannemann et al., 2008; Ongusaha et al., 2008; Wang et al., 2009]. Several mitogen-activated protein kinases involved in the MAPK pathway display changes in their activity, e.g. MAPKAPK2, MAPK3, MAPK14, suggesting that the MAPK pathway may also undergo regulation [Dhillon et al., 2007]. The estimated activities provide insight into the possible regulatory protein kinases/phosphatases responsible for mediating the phosphorylation changes.

**Figure 4.6:** Each kinase/phosphatase-substrate interaction is represented with three nodes: the source node represents the regulatory kinase or phosphatase, the middle node the targeted phosphosite and the sink node represents the substrate protein containing the targeted phosphosite. Edges are directed, to represent the regulatory interaction, and weighted. The edge linking the kinase/phosphatase to the phosphosite is weighted using the estimated activity of the regulator and the edge linking the phosphosite to the substrate protein is weighted with the phosphosite fold-change. Weights are transformed using the inverse of an empirical distribution function.
Subsequently, I resorted to the same comprehensive kinase/phosphatase substrate interaction network presented before to analyse possible functional interactions. Since each phosphosite may have a distinct functional impact these were represented as an independent node in the network (Figure 4.6). Associations between kinases/phosphatases and the substrate are represented with three different nodes, i.e. kinase/phosphatase, phosphosite and substrate protein, and two directed and weighted edges, i.e. one linking the kinase/phosphatase to the target phosphosite, and another linking the phosphosite to the substrate protein. This way a regulator may be associated with multiple phosphosites, and multiple phosphosites can be associated with the same substrate. This network provides a generic kinase/phosphatase-substrate interaction network. Hence, to define a context-specific network I overlapped the phosphorylation fold-changes and the kinases/phosphatases activity scores in the network (Figure 4.6). Edges linking kinases/phosphatases to their target phosphosites were weighted with the activity score, while edges linking the phosphosites to the substrate protein were weighted with the phosphorylation fold-changes. Both activity scores and fold-changes were separately transformed with the inverse of an empirical distribution function, thus highly changing phosphosites or kinases/phosphatases have a weight close to zero, and those not changing significantly are represented with a weight close to one. For interactions with no fold-changes or activity scores a weight of one was assigned. An important benefit of this representation is that it allows to efficiently transverse the network with weighted shortest paths (see section A.2.3). Therefore, a shortest path between two proteins is constituted by interactions supported by experimental evidence and by context-specific measurements. Shortest-paths will represent the path that minimised not only the number of interactions but also the sum of the edge weights. Some studies have shown that shortest paths and price-collection steiner trees reveal biologically relevant components of signalling pathways [Huang and Fraenkel, 2009; Tuncbag et al., 2012]. Steiner trees, or prize collecting steiner trees, is a computational problem that was first developed to infer from a network the minimum tree that maximizes the prices, that are represented as node weights. We explored that assumption to search for context-specific regulatory kinases and phosphatases upstream of differentially phosphorylated phosphosites in metabolic enzymes.

One of the most significantly changing metabolic reaction is pyruvate dehydrogenase (R_PDHm) (FDR=1.7e-197) (Figure 4.4 A, B). This is a metabolic
reaction converting pyruvate into acetyl-CoA catalysed by a protein complex involving pyruvate dehydrogenase enzymes (PDHA). This complex was associated with a significant increase in phosphorylation on the serine residue in position 232 of PDHA1 (PDHA1_S232) (FDR=1.45e-4) (Figures 4.2 C, Figure 4.4 A). Increased phosphorylation in any of the serines residues in positions 232, 293 and 300 have been shown to inactivate the enzyme and its function is only restored when all the three residues have been dephosphorylated [Korotchkina and Patel, 2001; Kato et al., 2008; Seifert et al., 2007]. This corroborates our hypothesis that PDHA complex activity is regulated by a phosphorylation change of PDHA1_S232, where increased phosphorylation in the KO cells is associated with a decreased metabolic flux. Also, decreased activity of PDHA1 in the KO cells suggests decreased intake of the glycolytic flux into the mitochondria and likely diverting it into the production of lactate, consistently with the Warburg effect and the CORE experiments.

**Figure 4.7:** Possible regulatory mechanism of PDHA1 S232 by ABL1 kinase. A) Kinase/phosphatase network upstream of PDHA1 S232 using manually curated interactions. Dark blue nodes represent kinases/phosphatases for which it was possible to estimate their activity using the phosphoproteomics measurements and their known targets. B) GSEA activity scores for the kinases/phosphatases involved in the network one and two levels upstream of PDHA1 S232. C) Schematic representation of the possible regulatory signalling mechanism controlling energy metabolism.

Having supported the importance of S232 phosphorylation status for the activity of PDHA1 I then explored possible regulatory kinases/phosphatases of this phosphosite. To this end, I used the previously assembled network and identified all the kinases and phosphatases upstream and calculated the weights of all the paths between them and PDHA1_S232. The number of upstream levels taken in consideration for this analysis was limited to two. While one level upstream reveals direct interactions, two steps upstream already increases significantly the
complexity of the search due to the complexity and high connectivity of signalling pathways [Jørgensen and Linding, 2010]. Hence, the associations immediately upstream are those for which we have higher confidence. One step upstream of PDHA1_S232 in the network shows PDK and PDPK kinases and PDP phosphatases, all well known regulatory proteins of PDHA1 (Figure 4.7 A). PDK, PDPK and PDP kinases/phosphatases have a similar activity score because the activity estimations are based only in one phosphosite, which is PDHA1_S232. Two steps upstream shows already nine possible regulators (Figure 4.7 A), supporting the previous affirmation that each step further upstream of the phosphorylation-site represents a steep increase in the number of possible associations. For only three of these kinases it was possible to estimate the activity, SRC, ABL1 and PRKCA, since the others had no measured substrates (Figure 4.7 A, B). Among all the possible paths of the upstream kinases the weighted shortest-path suggests ABL1 as the most actively changing kinase regulating the phosphorylation of Y243 in PDK1. Reassuringly, ABL1 has been shown to phosphorylate PDK1 and phosphorylation of PDK1_Y243 strongly increases the activity of the kinase [Hitosugi et al., 2011]. The second alternative mechanism for regulating PDHA1 activity would be via SRC that regulates the phosphorylation of multiple phosphosites of PDPK1, which in turn regulates also PDHA1_S232. Nevertheless, this option is less likely since SRC display decreased activity in KO cells. It is also important to note that phosphorylation may be a consequence of total protein changes and HIF has been associated with transcript regulation of PDHA1 [Baldewijns et al., 2010]. Although, this is a plausible justification, no measurement is available in the proteomics for PDHA1 and the transcript showed very low decrease in the expression, log₂ fold-change of -0.3, therefore strengthening the hypothesis that this may be a signalling regulatory event.

4.3 Discussion

Signalling and metabolism deregulation are among the hallmarks of cancer [Hana-han and Weinberg, 2000, 2011]. Understanding how cancer cells mediate their adaptation is a challenging task that requires an integrative perspective of the different biochemical processes of a cell. Advances in high-throughput experiments are increasing our capacity to acquire large-scale data-sets in different biological
processes and across hundreds of conditions [Mertins et al., 2016; Hakimi et al., 2016]. Therefore, developing mathematical methods capable of analysing and integrating the different biological measurements will provide useful insights into the regulatory interactions mediating the cellular adaptation.

HLRCC tumours display a characteristic mutation in the mitochondrial enzyme FH. This mutation leads to the loss of function of FH and requires tumour cells to adapt their metabolism, in particular, the CCM. To study the molecular adaption of the these tumours we have used tumour derived cell lines with a FH knockout and compared it to the same cell lines with FH reconstituted. In this chapter, I analyzed data from our collaborators characterizing the proteome, phosphoproteome and metabolome of these cell lines. I modelled the intracellular flux changes using quantitative experimental data measuring the consumption and secretion (CORE) rates of metabolites. Quantification of the metabolic rates, mmol/gDW/h, was important to discard possible confounding effects, such as cell number, and to robustly estimate the differences between the two cell lines. These revealed a significant increase in lactate secretion of the KO cell lines concordantly with the Warburg effect. The phosphoproteomics data-set also revealed strong changes involving several metabolic enzymes (Figure 4.1 C). Considering the implications of FH mutation in the phosphoproteome (Figure 4.1 C), it is fair to expect that both biological processes co-mediate their responses. Hence, this work provides for the first time a comprehensive characterisation of the phosphoproteome of HLRCC cell lines and its regulatory implications in the metabolic phenotype. The proteomics analysis revealed a strong concentration increase of VIM, an important gene involved in the EMT phenotype, and thereby corroborating the findings of previous transcriptomics data-set [Frezza et al., 2011] (Figure 4.1 B).

The phosphoproteomics and metabolomics experiments provided the base to develop a workflow to identify significantly changing phosphosites in metabolic enzymes and estimate their functional impact using intracellular flux changes (Figure 4.1 A). This approach allowed to systematically identify pairs of phosphosites and metabolic reactions, both mapping to the same enzyme, that are significantly changing in phosphorylation and in activity, thereby providing a way to identify putative regulatory phosphosites (Figure 4.4 A). I then expanded this approach to identify the possible regulators of the phosphorylation changes in the enzymes by using a comprehensive kinase/phosphatase-substrate interac-
tion network. Of note, the network representation was adapted to consider each phosphosite as a separate node and therefore enabling to assess their importance to the activity of the enzyme independently. This is particularly relevant since phosphorylation-sites may have a distinct functional impact in the enzymatic activity. A context-specific network was then built by weighting the edges with the phosphosites fold-changes and the estimated kinases/phosphatases activities. Normalising the weights enabled me to use classical methods to transverse the network, such as weighted shortest paths methods. This provides a computationally inexpensive way of obtaining paths between proteins that convey nodes changing significantly in the data-sets and edges that represent curated interactions.

I used this approach to identify enzymes undergoing phosphorylation regulation in HLRCC. Pyruvate Dehydrogenase (PDHA1) displayed one of the strongest significant change in the metabolic flux and it was associated with significant changes in phosphorylation in S232. In particular, it displayed decreased flux and increased phosphorylation in the KO cells. This result is supported by previous works that showed that increased phosphorylation of S232 reduces protein activity [Korotchkina and Patel, 2001; Seifert et al., 2007; Kato et al., 2008]. Therefore, this is a possible explanation for the decrease in the metabolic flux where S232 is the regulatory residue undergoing adaptation (Figure 4.4 A). Moreover, I hypothesised that ABL1 is the regulator of these changes as it shows the shortest weighted path with PDHA1_S232, compared to all the kinases two levels upstream in the signalling network (Figure 4.5 A, B). This regulation is mediated via PDK1 with ABL1 regulating the residue Y243 (Figure 4.5 A). In spite of the fact that PDK Y243 phosphosite is not measured it is brought into context by the comprehensive kinase/phosphatase-substrate network, and its importance is corroborated by previous studies assessing that increased phosphorylation leads to increased protein activity [Hitosugi et al., 2011]. Therefore, biological evidence seems to support that ABL1 when active may regulate the phosphorylation status of PDHA1, via PDK1, and therefore control its activity and the metabolic rewiring of energy production metabolism. This approach takes advantage of tailored methods to model signaling and metabolism and integrate them to systematically explore phosphorylation-sites that are relevant for regulating the activity of metabolic enzymes.
CHAPTER 5

SYSTEMATIC CHARACTERISATION OF THE PROTEOMICS AND PHOSPHOPROTEOMICS INTERACTOME IN COLORECTAL CANCER

The work presented in this chapter is part of a collaborative effort exploring the proteome and phosphoproteome of colorectal cancer cell lines. The results are from my own work and ideas, with discussion with Theodoros Roumeliotis, Steven Williams, Jyoti Choudhary and Ultan McDermott (Sanger Institute), input from Michael Schubert and Fatemeh Zamanzad and close supervision from Julio Saez-Rodriguez. All the experimental design and data acquisition were performed by Theodoros Roumeliotis and Jyoti Choudhary. This work is in validation phase and the subject of a paper in preparation.

5.1 Introduction

Cancer is one of the leading causes of deaths worldwide with an estimated 8.1 million deaths per year [Ferlay et al., 2014] and tackling it is one of the greatest challenges of our generations. Cancer is a complex disease and its phenotype arises from many different types of mutations and alterations. The landscape of cancer mutations is very heterogeneous among different tumours but also within
the same tumour type [Forbes et al., 2015; Stratton et al., 2009; Nik-Zainal et al., 2016; Martincorena and Campbell, 2015]. Hence, very few genes are consistently mutated across patients, as is the case of TP53, most of them are mutated only in a small percentage of the tumours [Cancer Genome Atlas Research Network et al., 2013; Forbes et al., 2015]. Extensive characterisation of the mutation landscape of tumours have provided unprecedented and vital information to understand the genetic rearrangement of tumours and how they compare to healthy cells from the same tissue [Cancer Genome Atlas Research Network et al., 2013; Martincorena and Campbell, 2015]. Recent large-scale efforts, such as the TCGA research network, provided comprehensive characterisation of many biological layers, e.g. DNA methylation, genomic or transcript expression, nevertheless the characterisation of the whole proteome, phosphoproteome and metabolome are still lagging behind [Zhang et al., 2014; Lawrence et al., 2015; Moghaddas Gholami et al., 2013; Hakimi et al., 2016]. While we are closer to a comprehensive picture of the mutation profile of tumours it has been shown that similar mutations across different tumour types display different functional outcome on protein activity. Therefore, similar to what has been done for genomic alterations, having a comprehensive molecular characterisation of other biological layers will help us understand what sets tumours with the same genomic background apart. This will provide valuable insights into the functional outcome of mutations and further understanding of the different tumour types and subtypes.

Semi-quantitative maps of the human proteome have become increasingly comprehensive over the last years [Wilhelm et al., 2014; Moghaddas Gholami et al., 2013; Uhlen et al., 2015; Zhang et al., 2014]. Proteomics and phosphoproteomics are arguably closer to the cell phenotype than transcript expression, e.g. protein measurements are currently used to classify subtypes of breast cancer according to different estrogen receptors. Cell lines have been used as representative models of the molecular profiles of tumors and valuable insights can be gained regarding the tumour response to drug treatments [Garnett et al., 2012; Iorio et al., 2016]. A recent study has acquired high coverage proteomics data for breast cancer cell lines and analysed the capacity to predict drug response [Lawrence et al., 2015]. With the technological developments is it fair to expect that more data-sets characterising the proteome of cancer cell lines will follow and thus shift the bottleneck from data generation to the development of methods capable of integrating these data-sets with the existing characterised biological layers.
Along these lines, I was involved in the analysis of proteomic and phosphoproteomic measurements from cell lines derived from bowel cancer patients. Bowel cancer is amongst the top four most predominantly causes of death in cancer worldwide, with colorectal cancer as the most common type [Ferlay et al., 2014]. Patients suffering from colorectal tumours have a poor prognosis and besides chemo and radiotherapies few drugs are available and have limited effectiveness. In this collaborative study we set to comprehensively characterise the proteome and phosphoproteome of 50 colorectal cancer cell lines. Specifically, I firstly analysed the reproducibility of the data-sets that we acquired and explored the implication of the proteome in the phosphoproteome. Secondly, I have explored the capacity of the proteomics to identify functional and direct protein-protein associations. Lastly, I developed a machine learning approach to identify mutations driving protein abundance changes and impairing protein-protein associations.

5.2 Results

5.2.1 Experimental setup and data normalisation

We have acquired high coverage proteomics and phosphoproteomics measurements across 50 colorectal cancer cell lines using isobaric labeling with tandem mass tags mass-spectrometry (TMT-MS) [McAlister et al., 2012; Rauniyar and Yates, 2014]. The cell lines were chosen from a comprehensive cancer cell line panel for which multiple molecular measurements are available, e.g. genomics, gene copy number variation, transcriptomics and drug response curves [Garnett et al., 2012; Iorio et al., 2016]. Proteomics and phosphoproteomics measurements were acquired in one initial 8-plex batch and then five 10-plex batches (Figure 5.1 A). TMT-MS requires data to be normalised within each batch to the same biological reference. Thus, cell line SW48 was picked as reference to reduce the genomic variability, since it was the cell line with the most stable genomic profile. Each peptide spectral mass (PSM) in each sample was log₂ transformed and then the batch reference cell line, SW48, was subtracted (Figure 5.1 B). The proteomics and phosphoproteomics measurements were grouped by protein and phosphorylation sites, respectively, by taking the median of all the PSM fold-changes. Spearman correlation of all pairwise combinations of the cell lines followed by unsupervised hierarchical clustering revealed a clear batch effect (Figure C.15 A,
C). Hence, to remove this batch effect for each protein and phosphosite a linear regression model was trained using the batch design as covariates and then removed from the original data-set (see section A.3.4). Reassuringly, re-applying the spearman correlation and hierarchical clustering did not reveal any evident batch structure (Figure 5.1 B, D).

While replicates were not generated for all cell lines, MDST8 cell line has two biological replicates one in batch 1 and another in batch 2. The replicates displayed strong reproducible results for both proteomics (spearman's rho \( r = 0.77 \), \( p \)-value \( \sim 0 \)) and phosphoproteomics (\( r = 0.61 \), \( p \)-value \( \sim 0 \)) (Figure 5.1 C, D). Additionally, proteomics measurements were acquired in an independent 10-plex batch following a different preparation protocol containing biological replicates of batch 6 (Figure 5.1 A). These biological replicates displayed an average correlation (mean spearman’s rho=0.73) significantly higher (t-statistic=17.96, \( p \)-value=5.1e-69) than the other non-replicate pairwise cell lines correlation (Figure 5.1 E). Due to significantly lower coverage this independent batch was not used for further downstream analysis. The same was applied to the MDST8 cell line measure in the batch 1 as it also displayed lower throughput than its replicate in batch 2. Cell line COLO-741 from batch 4 was wrongly annotated and was in fact a lung tissue cell line, hence it was also excluded (Figure 5.1 A).

The proteomics quantified 9,489 unique proteins with measurements in at least 50% of the cell lines, and 7,330 unique proteins measured across all cell lines (Figure 5.1 A). The phosphoproteomics covered 13,215 unique single phosphorylation sites (3,993 unique proteins) with measurements in at least half of the cell lines, and 3,801 unique phosphosites (1,887 unique proteins) measured across all cell lines. These data-sets provide rich and reproducible resources with increased coverage compared to existing cancer cell lines proteome studies [Moghaddas Gholami et al., 2013; Lawrence et al., 2015] and the characterisation of the phosphoproteome in a comprehensive set of colorectal cell lines.

\footnote{Due to the high correlation and number of measurements the module function provides a \( p \)-value of zero, considering that this is statistically not possible I abbreviated with \( \sim 0 \).}
Figure 5.1: Experimental overview and reproducibility analysis of the acquired data-sets. A) Experimental design of the 7 different batches acquired. The reference cell lines, i.e. SW48, in each batch is coloured in gray. Biological replicates are coloured in dark blue. A wrongly labeled lung cell line was discarded from any downstream analysis and colored in the diagram in red. B) Diagram of a representative 10-plex batch of TMT isobaric labeling and mass-spectrometry. Different colours represent the distinct isobaric labels used in each sample. C) Proteomics MDST8 biological duplicates correlation. D) Phosphoproteomics MDST8 biological duplicates correlation. E) Distributions of the spearman’s rho of batch 7 replicates compared to all the non replicate conditions correlation.
5.2.2 Proteome and phosphoproteome of colorectal cancer cell lines

Hierarchical clustering of the proteomics revealed a division of the colorectal cell lines into two different groups that showed extensive differences in the proteome (Figure 5.2 A). Unexpectedly, the phosphoproteomics data-set also displayed similar clustering to the proteomics. To quantify the similarity between the phosphosites measurements and its protein abundance I correlated all phosphosites in the phosphoproteomics data-set and the corresponding protein measured in the proteomics data-set. Only 1.87% (247) phosphorylation sites were not possible to correlate due to lack of proteomics measurements. While phosphosites and proteins measurements cover at least half of the cell lines the overlap between the two may be smaller. Therefore, to be consistent with the previous threshold, only correlations covering at least half of the cell lines were considered, leaving out 0.97% (125) of the phosphosites. Phosphosites showed a strong correlation with the protein abundance with a skewed normal distribution with a mean spearman correlation of 0.54, it also showed a significant change, Wilcoxon p-value $\sim 0$, when compared to randomised correlations (Figure C.16). Strong associations between protein abundance and phosphorylation status is expected in steady-state conditions as phosphorylation is naturally dependent of the protein abundance. Moreover, phosphoproteomics and proteomics were measured from the same sample preparation in order to minimise the technical variability and increase the similarity between the two experimental measurements. Consequently, to have a cell line clustering solely from the phosphoproteomics I removed the protein abundance from the phosphorylation levels of the sites by using linear regression with the protein abundance as a covariate (see section A.3.4). This allowed me to focus on the changes that happens solely due to phosphorylation. Subsequently, I redone the unsupervised hierarchical clustering on the phosphoproteomics normalised matrix and the obtained cell lines clustering was distinct from the previous and from the proteomics clustering (Figure 5.2 B, Figure C.17). Besides, there was no apparent clustering of the cell lines.

To analyse the different sources of variability in the protein abundance I performed Principal Component Analysis (PCA) on the fully measured proteome (Figure C.18 A). The first 10 principal components explained 53.59% of the total variance in the data-set with the first component capturing a small per-
Figure 5.2: Proteomics and phosphoproteomics cell lines correlation profiles. Cell lines were labeled according to their expression of HNF4A, consensus molecular subtype (CMS) and microsatellite instability in the heatmaps columns. Mutation profiles using recurring protein mutations were used to label the rows. Correlation has used as a distance metric for the hierarchical clustering. A) Proteomics hierarchical clustering of cell line pairwise spearman's rho. B) Phosphoproteomics normalised by protein abundance hierarchical clustering.
percentage (13.5%). Since transcription-factors have broad regulatory effects in the expression of the whole transcriptome and thereby they directly regulate protein abundance, I analyse the protein abundance of a comprehensive list of 1,444 human transcription-factors [Zhang et al., 2012, 2015]. I systematically correlated the transcription-factors abundance with principal component 1 (Figure C.18 B). The top correlated transcription-factor is HNF4A an important regulator of colorectal cancer cells [Chellappa et al., 2012; Zhang et al., 2014]. Moreover, HNF4A expression distinguishes the two different groups identified by the unsupervised clustering analysis, separating the colorectal cell lines into highly and lowly HNF4A expression (Figure 5.2 A). Other molecular characteristics were considered, i.e. consensus molecular subtypes [Guinney et al., 2015], microsatellite instability and different recurrently mutated genes, but none displayed any striking separation of the cell lines.

5.2.3 Estimating the activity of kinases and phosphatases

Next, I set to understand the regulatory proteins mediating the phosphorylation adaptation of the cell lines. Changes in the signaling pathways were measured by estimating the activity of kinases and phosphatases using the linear regression approach presented in Chapter 3. OmniPath [Türei et al., 2016], the same comprehensive resource of kinases/phosphatases-substrates interactions used in Chapter 4, was also used here (see section A.2.3). The high coverage of the phosphoproteomics data-set allowed me to estimate the activity of 218 kinases/phosphatases for the original phosphoproteomics and 216 kinases/phosphatases for the protein normalised phosphoproteomics.

The activity scores were used to analysed the downstream protein activity profiles of BRAF V600E and KRAS G12 mutations [Davies et al., 2002; Puxeddu et al., 2008]. V600E mutation renders BRAF kinase constitutively active in melanomas and thereby increases the phosphorylation and activity downstream [Davies et al., 2002]. Also several mutations leading to changes in the amino acid glycine in position 12 (G12) in KRAS were shown to be an important molecular alteration to predict the response to cetuximab by possibly rendering the kinase active [Engelman et al., 2008; Lièvre et al., 2008, 2006]. Therefore, I considered KRAS G12 amino acid modifications to alanine (A), cysteine (C), aspartic acid (D), serine (S) and valine (V) mutations. I explored the activity of BRAF,
KRAS and downstream protein kinases, i.e. PI3K, AKT, MEK and ERK (Figure 5.3 A, B). Kinases activities performed on the protein normalised and original phosphoproteomics data-sets did not show strong differences overall. MEK2 was the only reader that showed a stronger separation of the BRAF V600E mutants and the WT, and no significant distinction was visible in KRAS G12 mutations. Validation of the estimated kinases/phosphatases activities is a challenging task and it would require more tailored experiments, such as kinases knockouts. Nevertheless, these approaches have been applied successfully in different contexts [Casado et al., 2013; Mischnik et al., 2016; Ochoa et al., 2016]. The fact that there was a general lack of significant phosphorylation changes may highlight the challenge to interpret the signal transduction flow, which display dynamic regulatory mechanisms and may require a different experimental design than the one used for the proteomics measurements. For example, phosphoproteomics analysis of the cell lines would be more informative if the cells are first starved and then stimulated upon drug treatment.
5.2.4 Systematic identification of protein-protein interactions and protein complexes

Next, I set to evaluate the capacity of the proteomics data-set to extract functional associations between proteins. To this end, I performed all pairwise pearson correlations between the 7,330 proteins with measurements across all cell lines, generating a total of approximately 26 million unique correlations (see section A.3.6). Subsequently, I explored if highly correlated protein pairs represent functionally related protein interactions, thus I used STRING database [Jensen et al., 2009] as a resource of observed protein-protein interactions. STRING provides organism specific interactions and a confidence score that varies between 0 and 999, where the highest value represents the highest level of confidence and 0 the lowest. I filtered the data-set to consider only interactions found in human and with a confidence threshold of 900. Following this criteria 2.78% (237,606) of the total human interactions were used. Protein pairs reported in STRING displayed higher average spearman correlation coefficients than all the other pairwise combinations (STRING 0.32, all proteins 0.09) (Figure C.19). STRING ROC curve showed an area under the curve (AROC) of 0.75 (Figure 5.4 A) and the most significant correlating protein pairs were highly enriched for reported interactions, showing a low false positive rate for a true positive rate of approximately 40%. Furthermore, the capacity of our proteomics data-set was benchmarked against another proteomics and transcriptomics data-sets. Transcriptomics microarray experiments for 46 colorectal cell lines is available [Garnett et al., 2012], and fold-changes were calculated, similarly to the proteomics, against the SW48 cell line. An independent study acquired label-free proteomics for 20 breast cancer cell lines [Lawrence et al., 2015] measuring up to 1,388 proteins. Absolute log transformed intensities were used to estimate all pairwise protein correlations. Both proteomics data-sets displayed better capacity to extract known protein interactions than the transcriptomics (Figure 5.4 A).

While STRING recapitulates likely functional interactions I next evaluated the capacity of this approach to recover functional and direct protein associations. To this end, I used the CORUM database [Ruepp et al., 2008, 2010] which contains 1,331 protein complexes of which 84.15% (1,120) had at least 2 proteins measured in the proteomics data-set. Then I estimated the correlation of protein-pairs that are present together in at least one protein complex, and calculated
Figure 5.4: Inferring protein-protein interactions from proteomics measurements. A) ROC curve using protein-interactions from STRING as true positives and the protein pairwise spearman correlation \( p \)-values as the metric to set the thresholds to calculate the false positive and true positive rates. B) Similar analysis as in A) but true positives were inferred from CORUM protein complex data-base. C) Sample size effect on predicting protein-protein interactions was assessed by randomly sampled different number of samples. Groups of 10, 15, 20, 25, 30, 35, 40 and 45 samples were sampled 30 times each. ROC curve and AUCs were estimated as in A) and B). True positives are defined as the union from the STRING and CORUM data-sets. Randomised performance is estimated by shuffling the true positive labels and shown as light red. Dark red line present the obtained ROC and AUC using all samples. D) Same results as in C) but represented using boxplots.
a ROC curve to understand if there was an enrichment of interactions in highly correlated protein-pairs (Figure 5.4 B). Colorectal proteomics ROC displayed an AUC of 0.77, thereby showing a reasonable capacity to estimate functionally direct associations. The same analysis was performed with the previously presented proteomics and transcriptomics data-sets and similar results as before were obtained, with transcriptomics displaying lower capacity to predict reported protein functional interactions (Figure 5.4 B). The correlation distributions using the colorectal proteomics data-set showed a clear distinction between protein-pairs from the same complex and all the rest displaying a skewed bimodal distribution, highlighting different levels of agreement (Figure C.19).

Although, colorectal proteomics outperformed the breast cancer cell line panel one needs to consider that distinct technologies have been used to acquire the protein measurements, i.e. TMT-MS and label-free, and thereby technical reproducibility may be different as well as the coverage of the proteome. Besides, the number of measured cell lines in both studies is different, 50 colorectal and 20 breast cell lines. Therefore, to estimate the increased value of considering more cell lines to predict known protein-protein interactions I used our proteomics data-set and randomly sampled different numbers of cell lines. For each, randomly sampled set I performed the ROC curve analysis as before but now considering the union of the protein interactions between STRING and CORUM (Figure 5.4 C, D). Intuitively, an increasing number of cell lines displayed increased average predictive power, emphasising the importance of large cell line panels to recover reproducible and robust results. Of note, the $p$-values distributions of all the protein pair-wise spearman correlations did not display an uniform distribution, as it would be expected if all proteins were independent from each other (Figure C.19 A). This, re-emphasises that proteins in general are co-regulated.

### 5.2.5 Impact of mutations and phosphorylation in protein-protein associations

Considering the previous analysis it is reasonable to assume that strong correlations between protein pairs in the proteomics data-set is likely to point to a biologically direct and functional interaction. Protein-pairs significantly correlating may still display on specific cell lines a deviation from the linear association. These variable cases may point to protein modifications that affect their expres-
sion and possibly their association, thus I explored the possibility of mutations and PTMs being responsible for these variations. Mutations are often associated with impairment of protein expression by different types of mutations, e.g. frame-shift mutations, while phosphorylation can act as a marker for protein ubiquitination and thereby degradation, these events are termed phosphodegron [Swaney et al., 2013].

Figure 5.5: QQ-plots of the p-values obtained for the associations between the protein-protein residuals and the protein modifications. A) P-values of the associations between the protein-protein residuals and the genomic modifications. B) Same as in A) but for associations with phosphosites.

I analysed this hypothesis by focusing on 9,306 protein-pairs, which displayed a significant correlation (bonferroni adjusted p-value lower than 0.05) across the colorectal cell lines, thereby considering known as well as novel putative protein-protein interactions. A binary matrix identifying the genomic mutations for all the 49 cell lines [Forbes et al., 2015; Garnett et al., 2012; Iorio et al., 2016] was used rendering a total of 86,049 mutations, representing 2.09% of all the matrix. Presence of mutation was marked with a 1 and absence was marked with 0. The protein normalised phosphoproteomics data-set was binarized where all absolute fold-changes higher or equal than 1 was marked as 1 and the rest as 0. Binarization of the phosphoproteomics measurements is performed to guarantee that mutations and PTMs are comparable predictive features for the linear regression models used ahead (see section A.3.6). Phosphorylation events marked with 1
represented 2.52% (16,008) of the total matrix. To guarantee some degree of robustness in the results only mutations or PTMs that occur at least twice across the 49 cell lines are considered. This leaves 1.74% (1,446) unique mutations and 28.58% (3,708) unique phosphorylation-sites. For each protein pair all the single combinations with mutations and phosphosites were represented with a single feature linear regression (see section A.3.6). Accordingly, for each protein pair a mutation or phosphosite was used as a feature to predict the variation seen in the highly correlated proteins, using the correlation residuals between the two proteins as observations. This analysis allows me to search for mutations or phosphorylation events that are associated with the disagreement in specific cell lines between the protein pair. This can be used to systematically assign a putative effect of a protein modification in its abundance and possibly disrupt functional associations with other proteins by, for instance, impairing protein translation, affecting protein complex stability or marking the protein for degradation. The strength of the association between the protein-pair variation and the mutations or phosphosites were statistically assessed with a Wald test (see section A.3.6) and mutations provided stronger associations than phosphoproteomics (Figure 5.5).

Among the significant associations MSH2 and MSH6 (Figure 5.6 A) displayed a significant correlation and these are both involved in microsatellite instability phenotype an important aspect of colorectal cancer [Boland and Goel, 2010] (Figure 5.2 A). These proteins are involved in several protein complexes but they can also form a dimer, therefore it is reassuring that they display a significant agreement between their protein abundances profiles. A recurrent frameshift mutation in colorectal tumours in MSH6 [Forbes et al., 2015], F1088fs*5, is important to explain the variation seen on specific cell lines that display depleted protein abundance for both proteins (Figure 5.6 A). Of note, the mutation in the cell line HCT-116 did not seem to harbour any implication in the protein-interaction. The strongest association with mutations was the amino acid modification from an isoleucine to a phenylalanine in position 463, I463F, explaining the strong decrease in abundance of SEC13 and SEC13A (Figure 5.6 B). These proteins are involved in protein transport as part of the coat protein complex II (COPII) [Sato and Nakano, 2007; Tang et al., 1997]. Another example is the protein association between PIK3CB and PIK3R1 where a recurrent nonsense mutation R461* in PIK3R1 displays decreased protein abundance in itself but stronger alterations.
Figure 5.6: Representative significant associations between protein-protein residuals and protein modifications. Each plot displays the profiles of the two proteins involved in the interaction across the cell lines. Thick gray line represents the linear regression line and thinner gray lines the 0.5 and 1 range from the regression line. Cell lines displaying mutations in any of the two proteins are marked with a cross, cell lines with no mutations are represented with a circle. The modifications significantly associated with the residuals of the linear regression are marked with a red triangle.

in PIK3R1 (Figure 5.6 C). These proteins are involved in the signalling propagation by phosphorylation activating downstream AKT, and regulating several cellular processes involved in cell growth, survival and proliferation, to name a few [Miled et al., 2007; Jia et al., 2008; Wee et al., 2008]. Phosphoproteomics only displayed two significant associations (FDR <0.05), the strongest association is the previously reported [Mayya et al., 2009; Rigbolt et al., 2011] phosphoserine in the position 1,671 of the protein KMT2D where decreased phosphorylation is associated with decrease protein abundance (Figure 5.6 D). These are both histone methyltransferases involved in the regulation of tissue development [Lee et al., 2007; Cho et al., 2007; Lan et al., 2007]. These examples, provide evidence of the functional relationship between the protein pairs and of the usefulness of this approach to functionally annotate protein modifications. The number of possible associations is greatly affected by the number of samples considered. Lowly fre-
quent events can only be accurately measured and annotated with comprehensive analysis of increasing number of samples and across several tumour types.

5.2.6 Finding proteomics based predictions of drug response

An advantage of using cell lines is the capacity to systematically analyse the survival response to large panels of chemical compounds [Garnett et al., 2012; Iorio et al., 2016]. Cell lines panels are important exploratory tools for the identification of biomarkers that provide insight into the possible outcome of drug treatments [Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium, 2015; Iorio et al., 2016]. Thus, lastly I tried to identify associations with drug response using the proteomics, phosphoproteomics and kinases/phosphatases activities data-sets. Drug response was available for 47 out of 49 colorectal cell lines comprising 265 compounds [Garnett et al., 2012; Iorio et al., 2016] where 78.53% (9,781) of the matrix has AUC of the drug response curve. To estimate the capacity of the different data-sets to predict drug response I used each protein, phosphosite and kinases/phosphatase independently with single feature linear regression models where the observations are the drug response AUC values (see section A.3.6). There was a general lack of predictive power among the different data-sets (Figure 5.7 A, Figure C.20 A, C) where only proteomics displayed associations with an FDR lower than 5% (Figure 5.7 A, B). Among the top associated are some known multidrug resistant proteins, such as ABCB1, ABCB11, thus while reassuring these are poor follow-up targets [Krech et al., 2012; Katayama et al., 2014]. Phosphoproteomics and protein kinases/phosphatases activities displayed poor predictive power of drug response (Figure C.20 A, C) with associations displaying strong variation (Figure 5.6 B, D).

5.3 Discussion

Protein-proteins interactions are essential to the cell phenotype and are visible in many forms, for instance, transcription-factor binding, protein complexes and kinases/substrates associations. Recent technological advances in MS have allowed the acquisition of reproducible and high throughput proteomics and phosphopro-
teomics data-sets. That was corroborated with the data-sets presented in this chapter across 50 colorectal cell lines where biological replicates revealed strong correlations and great coverage of the phospho-proteome. The depth of the data-set was particularly noticeable considering that 7,330 proteins were consistently measured across all the different batches of cell lines. The proteomics profile of the cell lines recapitulated previous findings on transcriptomics data-sets associated with a clear division between the cell lines related with the expression of Hepatocyte nuclear factor 4-alpha (HNF4A) [Chellappa et al., 2012; Zhang et al., 2014] (Figure 5.2 A). HNF4A is a transcription factor thus is not surprising to see that it is responsible for a wide change of the proteome profile (Figure 5.2 A, Figure C.17). The phosphoproteomics data-set displayed strong correlations with the protein expression (Figure C.15). Biologically there is a dependence between protein phosphorylation and its abundance, e.g. protein degradation will naturally limit its phosphorylation and protein kinases/phosphatases display self-phosphorylating events. BRAF V600E mutant cell lines showed in general a marginal increase in the kinase activity of BRAF downstream target MEK2 (Figure 5.3 A, B). This lack of signal can be attributed to several factors, to enumerate a few: (i) fold-changes compared to a cancer cell line, i.e. SW48 cell lines, may not be the most appropriate control to visualise the signalling adaptation at a steady-state level using phosphoproteomics experiments, which arguably require higher sensitivity than proteomics, alternatives would be comparisons to pertur-
bations, i.e. drug treatments, or designing experiments with healthy tumours as a control; (ii) while BRAF V600E mutations render melanomas more sensitive to BRAF inhibitors, it may display different effects in colorectal cancers and drug response data to BRAF inhibitors do seem to suggest that with a compensatory negative feedback loop [Flaherty et al., 2012; Davies et al., 2002; Klinger et al., 2013].

The high coverage and the comprehensive number of cell lines measured allowed me to perform a broad analysis of protein-protein interactions. Proteomics data showed that it can recapitulate accurately and with precision (AROC 0.75) known regulatory protein-protein interactions reported in STRING [Jensen et al., 2009] (Figure 5.4 A). Moreover, it could identify protein-protein interactions of proteins involved in the same complexes (AROC 0.77). Taken together, these findings suggest that this current data-set can be a useful resource to identify putative functional and direct protein-protein associations. Moreover, the fact that both colorectal and breast cancer cell lines proteomics performed better than transcriptomics microarray (Figure 5.4 A, B) highlights the importance and the complementary value of protein measurements to elucidate different types of biological events. While microarray is a well established technology one could expect that more recent and accurate approaches to measure the transcriptome, i.e. RNA-seq, could possibly provide better predictive power for protein interactions. It is also interesting to note that a decrease in the number of cell lines in the proteomics data-set represented a decrease in the capacity to identify protein-protein interactions. This confirms that comprehensive cell lines and tumours panels are necessary to retrieve robust molecular associations. Besides, it is reasonable to assume that using samples from different tissues of origin will provide further increased predictive power, as it will cover tissue specific interactions.

The cellular phenotype is composed by many different regulatory processes that combine events at the genomic, transcriptomic, proteomic and phosphoproteomic level. Bearing this in mind, I have assessed the regulatory contribution of mutations and phosphorylation to the variability in protein interactions. Firstly, from all the 26 million pairwise correlations I selected those that correlate significantly and considered those as context-specific functional interactions. Secondly, from the functional interactions considered before I overlapped the mutations and phosphoproteomics modifications to explain the variability seen in the each highly correlated protein pair. Considering that protein interactions display
strong correlations the specific cell lines where they are dissimilar may point to cases where the interaction between the two is impaired. Thus, protein mutations or PTMs that are associated with this mismatch may be the regulatory event behind these alterations. Mutations displayed stronger associations than the phosphoproteomics measurements (Figure 5.5). Phosphorylation changes that regulate protein abundance, e.g. phosphodegron [Swaney et al., 2013], are arguably less frequent than mutations, specifically if we consider that cancer cell lines display very unstable DNA repair mechanisms. Moreover, genomic mutations and phosphorylation changes are only two out of many biological regulatory mechanisms affecting protein abundance and protein interactions. Currently, inferring the regulatory role of mutations and PTMs is still a challenge [Lemeer and Heck, 2009; Hornbeck et al., 2015] and methods that provide possible information of their functional impact have a wide utility. Hence, this analysis can be used as a resource to systematically study at a genome-scale the implication of mutations and phosphorylation changes in the proteome and in its interactome in colorectal cancer.

Many of these molecular variations may have a direct implication in the phenotype of the cell and this may provide different responses to drug treatments, for example, the over activation of growth inducing pathways may render cells more sensitive to growth inhibiting compounds [Sosman et al., 2012; Flaherty et al., 2012]. The identification of biomarkers is a current goal in cancer research that aims to identify changes in molecules that are predictive of the drug treatment outcome. Such markers provide vital information of how the treatment of a patient should be adapted to its own biological characteristics to grant maximum effectiveness to the treatment. Taking advantage of the availability of drug response curves for most of the analysed cell lines across 265 compounds, I used proteomics, phosphoproteomics and estimated kinases/phosphatases activities to build predictive models of drug response. Only the proteomics provided significant associations and most of the top associations were related with multidrug resistant proteins (Figure 5.7). More systematic analysis across many different types of cancer cell lines are being produced [Garnett et al., 2012; Iorio et al., 2016] considering multiple types of omics data, and these exploratory analysis are valuable resources for the identification of biomarkers. The complementary nature of the proteomics to the existing transcriptomics can be an added value to such exploratory studies, nevertheless from the results presented in this chapter
(Figure 5.7) one may consider that these can be marginal.
CHAPTER 6

CONCLUSIONS

In this thesis, I focused on the general and current challenge in systems biology of integrating and interpreting different biological measurements. Specifically, I devoted this thesis to the development of methods capable of modelling the interface between signalling and metabolism. In Chapter 1, I introduced the importance of having an integrated perspective of the many different molecular events occurring in a cell, and I illustrated it with several examples where this coordinated response is essential to understand the cellular phenotype. Computational methods have an increasing importance to robustly analyse biological experiments, therefore I introduced the current state-of-the-art approaches to model regulatory interactions. I then elaborated on the current advances, from small ODE models to large-scale networks, by building methods to systematically identify putative regulatory associations between metabolism and signalling. I also emphasised the complementary nature of the different biological processes and how despite being very different their intertwined response is vital for the precise and timely adaptation of the cell to numerous conditions.

In Chapter 2, I assessed the functional impact of phosphorylation in regulating the activity of metabolic enzymes. To this end, I acquired time resolved metabolomics measurements in yeast under salt and pheromone stimulations and paired it with an existing phosphoproteomics data-set. To identify regulatory phosphorylation-sites in metabolic enzymes I trained multilinear regression models using phosphoproteomics to estimate the metabolic changes. This presented a systematic and statistically robust approach that identified known and novel
regulatory phosphorylation-sites. This analysis also highlighted the importance of having high resolution of the initial response to the stimulus in order to capture the quick response of both signalling and metabolism. In Chapter 3, I went on exploring these data-sets further together with other published data-sets comprising a comprehensive panel of kinases/phosphatases knockouts and dynamic measurements upon perturbation of nitrogen metabolism and TOR1 signalling. Together, these were then used to infer transcriptional and post-transcriptional regulatory interactions of metabolic pathways. The results showed that steady-state experiments have strong confounding effects, such as cell cycle, that hinder the capacity of identifying direct associations between regulatory proteins, i.e. kinases, phosphatases and transcription-factors, and metabolic enzymes. In contrast, time resolved data-sets showed better predictive power of the metabolic changes. Machine learning models require large number of samples to obtain robust predictive models, therefore increasing availability of dynamic multi-omics data-sets across different conditions will provide stronger associations and allow to capture less frequent regulatory events.

In Chapter 4, I presented a novel approach to study the signal transduction and metabolic adaptations of hereditary leiomyomatosis and renal cell cancer cell lines upon restoration of FH expression. I resorted to a human genome-scale model to identify putative regulatory phosphosites in metabolic enzymes integrating novel phosphoproteomics and metabolomics data-sets. Specifically, the genome-scale model together with the consumption/secretion rates of metabolites were used to model the activity of metabolic enzymes using a modelling sampling approach. Consequently, the metabolic enzymes activity were associated with phosphorylation changes in their residues. Furthermore, I explored possible regulators of the phosphorylation changes in the metabolic enzymes using a comprehensive kinase/phosphatase-substrate interaction network. This novel approach takes advantage of existing tailored methods to model signaling and metabolism and interfaces them to study possible regulatory interactions between the two biological processes. A current limitation to this approach is that it relies solely on the comparison between two conditions, thus an important extension would be to integrate multiple conditions in other to extract more robust associations. Also, acquiring metabolite consumption/release rates is a laborious, time-consuming and relatively low-throughput experiment thus it may hamper the extension of this approach to larger sample sizes.
Lastly, in Chapter 5 I presented an approach to integrate genomics, phosphoproteomics and proteomics across a large panel of colorectal cancer cell lines. This approach was used to systematically identify protein modifications that affect abundance and potentially inhibit functional interactions, such as, the formation of protein complexes. Additionally, this analysis showed that MS based proteomics can accurately recover known protein-protein functional and direct interactions. The phosphoproteomics data-set showed in general a strong similarity to the proteomics measurements and also a lack of associations with other types of biological measurements. This, together with the poor predictive power demonstrated by the steady-state phosphoproteomics data-set presented in Chapter 3, emphasises the increased complexity of analysing this type of measurements compared to proteomics.

The necessity of systematic and comprehensive analysis to understand many biological questions lead to the usage of many different computational mathematical models [Kitano, 2002]. As the knowledge about the biological systems is only partial the computational models are always limited to have a simplified representation. Biological interactions are often modeled as simple linear associations to circumvent the usage of more complex mechanisms that require deeper knowledge and are generally computational intractable for large-scale systems. For these reasons the usage of ordinary least squares models (OLS) is widespread in the analysis of biological data-sets at various stages from data processing to remove potential batches effects [Ritchie et al., 2015] to perform genome-wide association studies (GWAS) [Pasaniuc and Price, 2016]. These models provide simple and robust mathematical and statistical formulation. Moreover, the simplicity of the formulation makes them easily interpretable and computational very efficient capable of accounting for thousands to even millions of samples and features. Nevertheless, the simplicity also brings the disadvantage of not being able to completely capture non-linear associations which are present in many biological processes. Moreover, these type of approaches generally require large sets of samples to be able to robustly recover associations and this has proven to be the major limiting step to apply these models more broadly. In spite of potential limitations, these mathematical models already have shown great potential to disentangle relevant biological mechanisms that would not be possible otherwise [Buettnner et al., 2015; Stegle et al., 2012; Aben et al., 2016; Marbach et al., 2016; Califano et al., 2012]. Furthermore, the modularity of the formulation of these
regression models allows to almost effortlessly integrate different types of biological data in a straightforward manner. Hence, providing an enticing framework to explore the complexity of the biological regulatory systems.

In this thesis, I showed that multilinear regression models can be reliably applied to identify regulatory interactions between signaling and metabolism. I illustrated this by using linear models in different contexts and to address different biological questions: (i) to identify regulatory phosphorylation-sites of metabolic enzymes, (ii) to estimate the activity of transcription-factors, kinases and phosphatases, (iii) to explore regulatory associations between signalling proteins and metabolic enzymes, and (iv) identify genomic alterations that lead to the disassembly of protein complexes. I also showed how tailored modelling approaches for signalling and metabolism can be integrated to hypothesise the impact of global kinases in the regulation of metabolism. This approach was used to identify a possible signalling regulation of energy production metabolism via the regulatory association of ABL1 protein kinase and the metabolic enzyme PDHA1.

The analysis of the different biological questions and data-sets presented in this thesis made me aware of the importance of the interoperability among different biological processes in mediating cellular adaptations. In particular, it emphasised (i) the interdependence between all the biochemical entities that constitute the cellular space, (ii) the increased value of time-resolved experiments over steady-state conditions to study regulatory and functional changes in the phosphoproteome, and thereby avoiding confounding effects, such as, protein abundance changes, (iii) the importance of co-designing biological experiments and computational analysis that are consistent, to ensure that both are reliably integrated, (iv) the advantage of current developments in mass-spectrometry based proteomics using TMT and SWATH-MS approaches over the classical methods to obtain higher coverage and reproducibility, and lastly (iv) the importance of an integrated perspective of the different cellular processes that, contrary to the idea of the central dogma, these act more as a component rather than as layer, and the variety of regulatory interactions among them forfeits the conceptual hierarchical structure.

Methods capable of integrating different biological measurements are still scarce and in general focus on well characterised pathways. Integrated approaches will provide useful tools to address many unanswered questions, to name a few, to systematically assess the impact of a post-transcriptional and post-translation
modifications in the protein enzymatic activity, or the identification of biomarkers predictive of the patient outcome to specific treatments. The continuous technological advances will allow us to measure more readouts and more accurately, and ultimately address some of the standing challenges and ask more complex questions. In particular, the advent of single-cell approaches allows us to currently measure thousands of transcripts per cell, and current efforts are being taken to also measure proteomics and metabolomics. Hence, this will allow us to question how diverse is the interactome between gene regulation, signalling and metabolism at a single cell level. Considering the challenges and the opportunities this is a very exciting time to be involved in the discovery and understanding of the cellular phenotypes and great breakthroughs can be expected.
A.1 Code dependencies and availability

All the computational analysis were performed in Python version 2.7.10 and are available under GNU General Public License V3 as GitHub projects in the following url https://github.com/saezlab. Plotting was done using Python modules Matplotlib version 1.4.3 [Hunter, 2007a] and Seaborn version 0.7.0 [Waskom et al., 2014]. Biological networks representation and analysis was implemented with Python module igraph version 0.7.1 [Csardi and Nepusz, 2006]. Generalised linear models were built using Python modules Sklearn version 0.17.1 [Pedregosa et al., 2011] and Statsmodels version 0.6.1 [noa, 2009]. Python modules Scipy version 0.17.1 [Jones et al., 2016] and Numpy version 1.11.1 [der Walt et al., 2011] were used to perform efficient numerical calculations and statistical analysis. Biological data analysis and structuring was carried out using Python module Pandas version 0.18.1 [McKinney and Others, 2010]. Differential expression analysis in R was performed under version 3.2.2 and using R package Limma version 3.30.0 [Ritchie et al., 2015].
A.2 Regulatory networks

A.2.1 Yeast kinase/phosphatase network

In Chapter 3, kinase/phosphatase-substrate network was assembled from PhosphoGrid data-base [Sadowski et al., 2013].

A.2.2 Yeast transcription-factor regulatory network

Transcription-factor regulatory network was assembled by Omar Wagih

In Chapter 3, specificities for a total of 177 transcription factors were collected in form of a position weight matrices (PWMs) from JASPAR [Mathelier et al., 2014]. Weight matrices were trimmed to remove consecutive stretches of low information content (<0.2) on either end. The log-scoring scheme defined in [Wasserman and Sandelin, 2004], was used to score potential target sequences against weight matrices. The log score is normalised to the best and worst matching sequence to the weight matrix, resulting in a value that lies between 0 and 1, where 1 denotes strong binding to the matrix and 0 denotes no binding. Genome wide gene expression profiles for 837 gene-knockout strains were collected from three studies [Kemmeren et al., 2014; Hu et al., 2007; Chua et al., 2006], of which 148/837 were a known transcription factor with a defined specificity weight matrix. Studies provided either a Z-score or $p$-value for each gene as a measure of over or under-expression, relative to the distribution of values for all genes. Two-tailed $p$-values were computed from Z-scores when a $p$-value was not provided [Chua et al., 2006]. In cases where TF knockout was repeated between studies, the lowest $p$-value for each gene was used. ChIP-ChIP tracks for 355 proteins were collected from four studies [Harbison et al., 2004; Rhee and Pugh, 2011; Tachibana et al., 2005; Venters et al., 2011], via the Saccharomyces genome database [Christie et al., 2004]. 144/355 of proteins were transcription factors with a defined specificity weight matrix. The TF-gene network was then defined as all TF-gene pairs with a $p$-value below 0.01 and contained a ChIP-ChIP region upstream of the regulated gene, which scored highly against the weight matrix of the TF (normalised logscore >0.9).
A.2.3 Human kinase/phosphatase network

In Chapters 4 and 5, the kinase/phosphatase-substrate network was assembled using OmniPath (www.omnipathdb.org) [Türei et al., 2016], a comprehensive resource assembling several databases of protein-protein interactions. Only manually curated phosphorylation interactions resources were used, i.e. PhosphositePlus [Hornbeck et al., 2015], PhosphoELM [Dinkel et al., 2011], DEPOD [Duan et al., 2015], HPRD [Keshava Prasad et al., 2009] and Signor [Perfetto et al., 2016].

To explore the regulatory kinases/phosphatases upstream a given phosphosite the network is then adapted to represent each unique phosphosite as a node. The network is then weighted with phosphosite fold-changes and the kinases/phosphatases activities scores. Weights are calculated with the inverse of the empirical distribution function so that lighter edges represent more significantly changing events (Figure 4.6). The weights of the kinases/phosphatases are set to the edges linking them to the targeted phosphorylation-sites, and the weights of the phosphosites are set to the edges linking them to the substrate protein. The levels upstream of a specific phosphorylation-site is calculated by obtaining the inbound neighborhood of the node, then for each kinase/phosphatase the shortest path is calculated.

A.3 Linear regression models

Ordinary least squares models (OLS) were used throughout this thesis and therefore I collated the details of each particular analysis in the following subsections.

A.3.1 Identification of putative regulatory phosphosites in metabolic enzymes

In Chapter 2, only metabolites and phosphosites changing significantly in at least one time point were considered. A genome-scale metabolic reconstruction [Mo et al., 2009] was used to overlap the metabolomics and phosphoproteomics measurements. Highly connected metabolites and cofactors were removed from the network. The metabolic network was then used to identify the metabolic enzymes one and two reactions away from the significantly changing metabolites.
and to overlap the measure phosphosites of the enzymes. This mapping was then
used to build multilinear regression models:

\[ Y = \beta X + \psi \]  \hspace{1cm} (A.1)

Where, the observed variable \( Y \) represents the metabolite fold-changes and \( X \)
the phosphosites of the surrounding enzymes both variables measured across the
time-points in the different conditions (Figure 2.5 A). The predictive power of the
linear model was evaluated using the coefficient of determination. Considering
that each metabolite has a different number of surrounding phosphosites, to have
comparable measures an adjusted coefficient of determination was used. This
takes in consideration the number of features of the model:

\[
Adjusted R^2 = R^2 - \left(1 - R^2\right) \frac{P}{N - P - 1}
\]  \hspace{1cm} (A.2)

Where, \( R^2 \) represents the coefficient of determination, \( N \) the number of ob-
servations and \( P \) the number of features. Features, i.e. phosphosites, importance
was estimated by the \( \beta \) values of the linear regressions (A.1).

**A.3.2 Estimating metabolomics changes using transcription-factor and kinases/phosphatases activities**

In Chapter 3, linear models with combined L1 and L2 regularization, Elastic
Net, were used. Elastic net regularization simplifies the complexity of the model
by removing the least important features, similar to Lasso regularisation, but
also considering a L2 regularization, similar to Ridge, to avoid random feature
elimination when collinearity exists among the features.

To infer the predictive power within each data-set across all the measured
ions (Figure 3.3 B), for each metabolomics data-set, ions displaying low vari-
ation across the samples were discarded by considering only those that showed a
standard deviation higher than 0.4. K/P activities were filtered to only consider
kinases or phosphatases with an activity score estimated in at least 75% of the
samples of each data-set, the remaining missing values were replaced with zeros
for the machine learning approaches. For the Elastic net regressions different
alphas were tested and an alpha of 0.01 obtained the best overall performance,
therefore this was used in all the models. For each metabolite a leave-one-out
cross-validation was used, thus all but one sample were used to train the linear regression model and then the test sample was used to estimate the metabolite fold-change. Performing this systematically across all metabolites and conditions generated a predicted matrix for which each value is estimated independently. The agreement between the measured and predicted ions fold-changes was calculated with Pearson correlation coefficients across rows (ions) and columns (conditions).

A.3.3 Identifying regulatory interactions between kinases, phosphatases and transcription-factors and metabolites

In Chapter 3, protein-metabolite associations were inferred only using the time-resolved metabolomics data-sets. For the TF activities only the nitrogen metabolism perturbations were used considering that no transcriptomics data was available for the NaCl/pheromone perturbations. As before, Section A.3.2, only TF and K/P profiles estimated in at least 75% of the conditions were considered.

The capacity of predicting each ion fold-change in each condition was tested using k-fold cross-validation, where each fold corresponds to the samples of each condition, thus 3 and 5 folds were used for the TF and K/P activities, respectively. The alpha parameter of the Elastic Net models was estimated using a bootstrap approach of ten iterations leaving out 20% of the samples. A range of 100 alphas was considered as default by the Sklearn python module [Pedregosa et al., 2011]. Train and test features, TFs and kinases/phosphatases activities, were standardised, and the observed variables, metabolomics, were centered before training the linear model. The predictive power of each ion in each condition was estimated by using the k-fold models, inferring the agreement between predicted and measured in the left-out condition using Pearson coefficient and coefficient of determination metrics. The top predicted ions were those that displayed a Pearson correlation p-value lower than 0.05 and an coefficient of determination higher than zero.

For the top predicted ions the feature importance was estimated using all the conditions together with two bootstraps. The first bootstrap was 20 iterations and leaves out 20% of the samples out, for each iteration an inner bootstrap with 10 iterations leaving out another 20% of the data is performed to estimate the alpha of the Elastic net. This estimates 20 coefficients for each feature-metabolite
association. As before, train features and observations are standardised and centered. The most important features per ion are estimated by taking the mean of the coefficients and the Pearson correlation between the protein activity and the ion fold-change.

A.3.4  Normalise phosphorylation with protein abundance

In Chapter 5, protein phosphorylation displayed strong correlation with protein abundance (Figure C.16). To assess the changes solely driven by phosphorylation, I normalised the phosphoproteomics measurements by protein abundance. To this end, I considered only phosphorylation-sites for which there was a matching protein abundance measurement, i.e. 14,117 (98.2%) of the measured sites. Then for each phosphosite I fitted a linear regression model with the protein abundance across the cell lines as a covariate and the dependent variable representing the phosphorylation measurements of the phosphosite also across the conditions. The model was also fitted with an intercept and a noise term:

\[ y = \beta.x + \psi \]  \hspace{1cm} (A.3)

Where \( y \) represents the phosphorylation measurements, \( x \) the proteomics, \( \psi \) the noise term and \( \beta \) the coefficient of the proteomics measurements. Once the coefficient of the covariate is estimated it is used to regress-out the protein abundance estimated effect on the phosphorylation by:

\[ y' = y - \beta.x \]  \hspace{1cm} (A.4)

Where, \( y' \) represents the protein normalised phosphorylation measurements.

A.3.5  Protein-pairs correlations and associations with protein modifications

In Chapter 5, for this analysis only proteins measured across all the cell lines were considered leaving 7,330 unique proteins. Protein correlation was estimated using the computationally efficient function \textit{spearmanr} from SciPy package [Jones et al., 2001; der Walt et al., 2011]. This allow me to estimate the correlation coefficient and \( p \)-value of approximately 26 million unique pairwise correlations.
A stringent threshold was used to select the significantly correlating protein pairs using bonferroni as a multiple hypothesis correction test. Only protein-pairs with an adjusted $p$-value lower than 0.05 were considered as a true positive interaction and used for the protein modifications analysis. QQ-plots were performed to compare the observed versus the theoretical estimated $p$-values following a randomly uniform distribution.

For each significantly correlated protein-pair the residuals of the correlation, i.e. the differences of the measurements from the fitted linear regression line, were estimated using the mutation and the phosphoproteomics changes. An initial linear regression similarly to A.3 between the two protein pairs is performed to assess how well the two proteins associate. The residuals of the fit is then used in a second linear regression as the dependent variable, the independent variables are the binary protein modification matrices, i.e. mutations and phosphorylation, where 1 indicates the presence of the modification and 0 the absence. Only mutations or phosphosites falling on the protein-pair are considered as features for the linear regression. Each feature is considered and trained separately. A univariate Wald statistical test, variation of the maximum likelihood of the associations beta, is calculated for each feature independently to estimate the significance of the protein modification on the fit of the linear model. $P$-values are then corrected for multiple hypothesis test using the Benjamini-Hochberg False Discovery Rate.

A.3.6 Drug response associations with proteomics, phosphoproteomics and protein activities

In Chapter 5, drug response associations were estimated similarly to the genomics and phosphoproteomics analysis, Section . Drug response curves area under the curve (AUC) were used as observations. Each feature data-set was analysed separately and for each feature/drug combination a linear regression was performed. The effect size of the association was statistically assessed using a Wald test, $p$-values were then corrected for multiple hypothesis using FDR.
A.4 Sampling intracellular metabolic flux distributions

In Chapter 4, the human metabolic reconstruction used was Recon1 [Duarte et al., 2007]. In the KO cell lines metabolic model FH catalysed reactions are inactivated by setting a lower and upper bound of zero. To constraint the metabolic model with the CORE experiments I first considered for intake only metabolites that are either measured or present in the constituents of the DMEM 4196 medium, all the other metabolites were constrained to an intake of zero. Oxygen consumption rates measured with Seahorse experiment was acquired and used to constrain the oxygen exchange reaction. Subsequently, I minimise the absolute differences between the measured and the predicted rates by performing a FBA where the objective function is the minimisation of these differences. This estimates the best possible fit to the experimental data considering the model and the selected metabolite list for intake. The fitted intake and secretion rates are then used to constrain the metabolic model with hard constraints in the exchange reactions bounds and a MCMC sampling approach, optGpSampler [Megchelenbrink et al., 2014], is used to obtain intracellular flux distributions for the KO and WT cell lines. To guarantee an uniform sampling of the solution space each condition was sampled 10,000 times and each sample was performed with 2,500 steps [Megchelenbrink et al., 2014].
APPENDIX B

EXPERIMENTAL METHODS AND MATERIALS

B.1 Yeast cells

B.1.1 Strain, growth and sample preparation

Experiments were performed with close supervision of Zrinka Nakic and Mattia Zampieri. Methods and materials were written with the help of Zrinka Nakic.

In Chapter 2, the Saccharomyces cerevisiae strain used for the salt and pheromone dynamic experiments was BY4741 as in Vaga et al. [Vaga et al., 2014]. This strain is provided with a Cdc28-as allele that can be directly inhibited by means of 1-NA-PP1, the ATP analog 8-P1 analog 8. Cells were grown in 500-ml shake flasks at 30°C in 50 ml SD medium to an OD600 of 0.6. The ATP analog was added to a final concentration of 10µM. One hour after Cdc28 inhibition cells were perturbed with NaCl to a final concentration of 0.4M, pheromone to a final concentration of 1 µM and NaCl and pheromone to a final concentration of 0.4M and 1 µM, respectively. Cells were extracted by vacuum-filtering culture aliquots on a 0.45 µm pore size nitrocellulose filter (Millipore). The filter was immediately transferred to 3 ml 2:2:1 MeOH/AcN/ddH2O precooled at -30°C. Samples for LC-MS/MS were supplemented with 200 µl uniformly-labeled 13C E. coli extract as internal standard and dried completely in a vacuum centrifuge.
The dried extracts were resuspended in 100 µl MilliQ water before analysis.

**B.1.2 Acquisition of intracellular metabolite levels and quality control**

Targeted metabolomics was performed by LC-MS/MS as described before [Buescher et al., 2010]. The mass-spectrometer was operated in negative mode. Data acquisition and peak integration were performed with the Xcalibur software version 2.07 SP1 (Thermo Fisher Scientific) and in-house integration software. Metabolite peak areas were normalized to uniformly-labeled 13C internal standards. Samples were supplemented with 54 carbon labeled extracts for which both the labeled and unlabeled spectral mass peaks were manually identified and curated. The manual quality assessment of the peaks lead us to retain 26 metabolites which could be reliably identified from the sample spectra.

Untargeted metabolomics was performed by direct flow double injection of extracts on an Agilent 6550 series quadrupole TOF MS operated in negative mode. In total 11,190 ions were detected across all samples. The detected ions were then mapped against an existing ion mass library of metabolites generated from the yeast genome-scale model iMM904 [Mo et al., 2009]. A stringent mass tolerance of 0.001 Da was used. 452 out of 647 ion entries in the library were identified in our data-set and these were kept for the downstream analysis. Considering that natural modifications can occur and change metabolite mass we used a rigorous annotation of the detected ions to only consider deprotonated metabolites, therefore reducing the number of detected and annotated ions to 196. These filtering steps allowed us to consider only highly confidently annotated ions which mapped to 270 yeast metabolites. High concentrations of salt in the NaCl perturbation experiment resulted in a strong effect on the ion matrix in the QTOF-MS measurements. To prevent this matrix effect from affecting data analysis we normalised the data to the second time-point (25 seconds) instead of the 0 seconds timepoint.

Statistical significance of the ions fold-changes for the QTOF-MS measurements were estimated with a two-sided t-test followed by multiple hypothesis correction with false discovery rate.
B.1.3 Compendium of high throughput publicly available yeast studies

In Chapter 3, the phosphoproteomics data-set [Vaga et al., 2014] and paired metabolomics were complemented with publicly available transcriptomics, phosphoproteomics and metabolomics data-sets. To this end, I considered a transcriptomics data-set measured across 1,484 deletion mutants [Kemmeren et al., 2014]. This data-set was complemented by phosphoproteomics [Bodenmiller et al., 2010] and metabolomics [Schulz et al., 2014] measurements covering a total of 125 K/P knockouts. Metabolomics and phosphoproteomics data-sets overlap in 115 knockout conditions, and metabolomics intersects the transcriptomics in 45 knockout conditions. Time-resolved experiments were also considered by using a perturbation study of the nitrogen metabolism and TOR signalling [Oliveira et al., 2015b,a]. Transcriptomics measurements covered 5,620 transcripts across all the time-points of the conditions. Phosphoproteomics captured the profile of 1,660 single phosphorylated phosphosites (84.8% serines, 14.2% threonines and 1.0% tyrosines) over the same time-points. Intracellular metabolomics were acquired with QTOF-MS and quantified a total of 146 ions, after quality filtering, across all conditions and time-points.

B.2 Human cells

B.2.1 Hereditary leiomyomatosis and renal cell cancer cell lines

Experiments were performed by Marco Sciacovelli and Sofia Costa. Note that all the below information was kindly provided by Marco Sciacovelli and Sofia Costa.

B.2.1.1 Growth and sample preparation

Human FH-mutant UOK262 and FH-reconstituted UOK262pFH cells were obtained as previously described [Frezza et al., 2011]. All cells were grown in DMEM (Gibco 41966-029) supplemented with 10% heat inactivated FBS (Gibco 10270-106).
B.2.1.2 Consumption and release quantification of metabolites

OK262 and UOK262pFH (1.5x10^5) were plated onto 6-well plates and allow to grow for 16h. 200 µl of cell culture media were collected from each well immediately after (t=0) and after additional 24 hours of incubation. Cells from the same plate were used for counting and then lysated in RIPA buffer for measurement of protein content. The collected media were centrifuged at 4°C for 10 min at max speed and 50 µl of the supernatant extracted in 750 µl of cold metabolites extraction buffer (MEB). The solution was centrifuged at 4°C for 10 min at max speed and the supernatant was transferred onto LC-MS vials for metabolomics analyses.

LC-MS analysis of sample extracts was performed on a QExactive Orbitrap mass spectrometer coupled to Dionex UltiMate 3000 Rapid Separation LC system (Thermo). The liquid chromatography system was fitted with a SeQuant ZIC-pHILIC (150mm 2.1mm, 5µm) with guard column (20mm 2.1mm, 5µm) from Merck (Darmstadt, Germany). The mobile phase was composed of 20mM ammonium carbonate and 0.1% ammonium hydroxide in water (solvent A), and acetonitrile (solvent B). The flow rate was set at 180µL x min-1 with the following gradient: 0 min 70% B, 1 min 70% B, 16 min 38% B, 16.5 min 70% B, hold at 70% B for 8.5 min. The mass spectrometer was operated in full MS and polarity switching mode. Medium from five independent cell cultures were analysed for each condition and samples were randomised in order to avoid bias in sample analyses due to machine drift. The acquired spectra were analysed using XCalibur Qual Browser and XCalibur Quan Browser software (Thermo Scientific) by referencing to an internal library of compounds.

Absolute quantification of metabolites in the cell culture medium was performed by interpolation of the corresponding standard curves obtained from commercially available compounds running with the same batch of samples. For each spent medium sample and each metabolite, the measured concentration spent was converted to consumption/release (CORE) data (molar amounts per dry weight per unit time) adapting the approach described in [Jain et al., 2012].

B.2.1.3 Mass-spectrometry based proteomics and phosphoproteomics

Following mass-spectrometry analysis was carried out in the laboratory of Pedro Cutillas. Note that all the below information was kindly provided by Pedro Cutil-
Proteomics experiments were performed using mass spectrometry as reported [Rajeeve et al., 2014; Casado et al., 2013]. Cells were lysed in urea lysis buffer (8 M urea, 10 mM Na3VO4, 100 mM β-glycerol phosphate and 25 mM Na2H2P2O7 and supplemented with phosphatase inhibitors (Sigma)) and proteins reduced and alkylated by sequential addition of 1 mM DTT and 5 mM iodoacetamide. Immobilized trypsin was then added to digest proteins into peptides. After overnight incubation with trypsin, peptides were desalted by solid phase extraction (SPE) using OASIS HLB columns (Waters) in a vacuum manifold following the manufacturers guidelines with the exception that the elution buffer contained 1 M glycolic acid. For phosphoproteomics analysis the peptides were enriched using TiO2 beads (GL Sciences) similarly to that described [Montoya et al., 2011; Casado et al., 2013] with minor modifications. Samples were analysed with LC-MS/MS using a LTQ-Orbitrap mass-spectrometer. Peptides were identified using Mascot against SwissProt human protein database and quantified applying Pescal as previously described [Casado et al., 2013].

Dried peptide extracts were dissolved in 0.1% TFA and analysed by nanoflow LCMS/MS in an LTQ-orbitrap as described [Rajeeve et al., 2014; Casado et al., 2013]. Gradient elution was from 2% to 35% buffer B in 90 min with buffer A being used to balance the mobile phase (buffer A was 0.1% formic acid in water and B was 0.1% formic acid in acetonitrile). MS/MS was acquired in multistage acquisition mode. MS raw files were converted into Mascot Generic Format using Mascot Distiller (version 1.2) and searched against the SwissProt database (version 2013.03) restricted to human entries using the Mascot search engine (version 2.38). Allowed mass windows were 10 ppm and 600 mmu for parent and fragment mass to charge values, respectively. Variable modifications included in searches were oxidation of methionine, pyro-glu (N-term) and phosphorylation of serine, threonine and tyrosine. Results were filtered to include those with a potential for false discovery rate less than 1% by comparing with searches against decoy databases. Quantification was performed by obtaining peak areas of extracted ion chromatographs (XICs) for the first three isotopes of each peptide ion using Pescal [Casado and Cutilias, 2011; Cutilias and Vanhaesebroeck, 2007]. To account for potential shifts in retention times, these were re-calculated for each peptide in each LCMS/MS run individually using linear regression based on com-
mon ions across runs (a script written in python 2.7 was used for this retention
time alignment step). The mass and retention time windows of XICs were 7 ppm
and 1.5 min, respectively.

B.2.2 Colorectal cancer cell lines

Experiments were performed by Theodoros Roumeliotis. Note that all the below
information was kindly provided by Theodoros Roumeliotis.

B.2.2.1 Cell culture and reagents

Cells were grown in either DMEM/F12 medium (Gibco) supplemented with 10%
fetal calf serum (v/v) (Gibco) and 50 U/ml penicillin, 50 mg/ml streptavidin
(Gibco), or RPMI 1640 medium (Gibco) supplemented with 10% fetal calf serum
(v/v) (Gibco), 50 U/ml penicillin, 50 mg/ml streptavidin (Gibco), 2.5 mg/ml
glucose (Sigma-Aldrich) and 1 mM sodium pyruvate (Gibco), and maintained at
37°C in a humidified atmosphere at 5% CO2. Cells were harvested by incubating
with TrypLE (Gibco) until detached, and washing twice with cold PBS solution
before snap freezing on dry ice.

B.2.2.2 Protein digestion and TMT labeling

The PBS washed cell pellets were dissolved in 150 µL 0.1 M triethylammonium
bicarbonate (TEAB), 0.1% SDS with pulsed probe sonication (EpiShear, power
40%) on ice for 20 sec and direct boiling at 95 °C in a preheated heat block for 10
min. The sonication-boiling procedure was performed twice and cellular debris
was removed by centrifugation at 12,000 rpm for 10 min. Protein concentration
was measured with Quick Start Bradford Protein Assay (Bio-Rad) according to
manufacturer’s instructions. Aliquots containing 100 µg of total protein were
prepared for trypsin digestion. Cysteine disulfide bonds were reduced with a
final concentration of 5 mM tris-2-carboxymethyl phosphine (TCEP) followed by
1 h incubation in heating block at 60 °C. Cysteine residues were blocked with a
final concentration of 10 mM freshly prepared Iodoacetamide (IAA) solution and
30 min incubation at room temperature in dark. Trypsin (Pierce, MS grade) was
added at mass ratio 1:30 for overnight digestion. The resultant peptides were
diluted up to 100 µL with 0.1 M TEAB buffer. A 41 µL volume of anhydrous
acetonitrile was added to each TMT 10-plex reagent (Thermo Scientific) vial and after vortex mixing the content of each TMT vial was transferred to each sample tube. The labeling reaction was quenched after 1 hour by the addition of 8 µL 5% hydroxylamine. Samples were combined and the mixture was dried with speedvac concentrator and stored at -20 °C until the high-pH Reverse Phase (RP) fractionation. The SW48 cell line was used as the reference sample to enable inter-experimental comparison and was newly cultured in parallel with the rest of the cell lines in each sample batch.

B.2.2.3 Peptide fractionation

High pH Reverse Phase (RP) peptide fractionation was performed with the Waters, XBridge C18 column (2.1 x 150 mm, 3.5 µm, 120 Å) on a Dionex Ultimate 3000 HPLC system equipped with autosampler. Mobile phase (A) was composed of 0.1% ammonium hydroxide and mobile phase (B) was composed of 100% acetonitrile, 0.1% ammonium hydroxide. The TMT labelled peptide mixture was reconstituted in 100 µL mobile phase (A), centrifuged and injected for fractionation. The multi-step gradient elution method at 0.2 mL/min was as follows: for 5 minutes isocratic at 5% (B), for 35 min gradient to 35% (B), for 5 min gradient to 80% (B), isocratic for 5 minutes and re-equilibration to 5% (B). Signal was recorded at 215 and 280 nm and fractions were collected in a time dependent manner every 30 sec. The collected fractions were dried with SpeedVac concentrator and stored at -20 °C until the LC-MS analysis. For the replication sample set peptide fractionation was performed on reversed-phase OASIS HLB cartridges at high pH and up to 10 fractions were collected for each set.

B.2.2.4 Phosphopeptide enrichment

The high-pH peptide fractions were reconstituted in 10 µL of 20% isopropanol, 0.5% formic acid binding solution and were loaded on 10 µL of phosphopeptide enrichment IMAC resin (PHOS-Select Iron Affinity Gel) already washed and conditioned with binding solution in custom made filter tips fitted on the eppendorf tubes caps. The resin was washed three times with 40 µL of binding solution and centrifugation at 300 g after 2 h of binding and the flow-through solutions were collected. Phosphopeptides were eluted three times with 70 µL of 40% acetonitrile, 400 mM ammonium hydroxide solution. Both the eluents and
flow-through solutions were dried in a speedvac and stored at -20 °C until the phosphoproteomic and proteomic LC-MS analysis respectively.

B.2.2.5 LC-MS Analysis

LC-MS analysis was performed on the Dionex Ultimate 3000 UHPLC system coupled with the Orbitrap Fusion Tribrid Mass Spectrometer (Thermo Scientific). Each peptide fraction was reconstituted in 40 µL 0.1% formic acid and a volume of 7 µL was loaded to the Acclaim PepMap 100, 100 µm 2 cm C18, 5 µm, 100 Å trapping column with the iPickUp mode at 10 µL/min flow rate. The sample was then subjected to a multi-step gradient elution on the Acclaim PepMap RSLC (75 µm 50 cm, 2 µm, 100 Å) C18 capillary column (Dionex) retrofitted to an electrospray emitter (New Objective, FS360-20-10-D-20) at 45 °C. Mobile phase (A) was composed of 0.1% formic acid and mobile phase (B) was composed of 80% acetonitrile, 0.1% formic acid. The gradient separation method at flow rate 300 nL/min was as follows: for 95 min gradient to 42% B, for 5 min up to 95% B, for 8 min isocratic at 95% B, re-equilibration to 5% B in 2 min, for 10 min isocratic at 5% B.

Precursors were selected with mass resolution of 120k, AGC 310⁵ and IT 100 ms in the top speed mode within 3 sec and were isolated for CID fragmentation with quadrupole isolation width 0.7 Th. Collision energy was set at 35% with AGC 110⁴ and IT 35 ms. MS3 quantification spectra were acquired with further HCD fragmentation of the top 10 most abundant CID fragments isolated with Synchronous Precursor Selection (SPS) excluding neutral losses of maximum m/z 30. Quadrupole isolation width was set at 0.5 Th, collision energy was applied at 45% and the AGC setting was at 610⁴ with 100 ms IT. The HCD MS3 spectra were acquired within 120-140 m/z with 60k resolution. Targeted precursors were dynamically excluded for further isolation and activation for 45 seconds with 7 ppm mass tolerance. Phosphopeptide samples were analyzed with CID-HCD method at the MS2 level. MS level AGC was set at 610⁵, IT was set at 150 ms and exclusion duration at 30 sec. AGC settings for CID and HCD fragmentation were 510⁴ and 210⁵ respectively. The fractions for the replication and CRISPR/cas9 sets were analysed with 180 min and 300 min LC-MS runs respectively and the analysis was repeated by setting an upper intensity threshold at 2-510⁶ to capture lower abundant peptides. The total data collection was accomplished with 475
LC-MS runs in 950 hours of analysis.

**B.2.2.6 Protein identification and quantification**

The acquired mass spectra were submitted to SequestHT search in Proteome Discoverer 1.4 for protein identification and quantification. The precursor mass tolerance was set at 20 ppm and the fragment ion mass tolerance was set at 0.5 Da for the CID and at 0.02 Da for the HCD spectra used for the phosphopeptide analysis. Spectra were searched for fully tryptic peptides with maximum 2 miss-cleavages and minimum length of 6 amino acids. TMT6plex at N-termimus, K and Carbamidomethyl at C were defined as static modifications. Dynamic modifications included oxidation of M and Deamidation of N,Q. Maximum two different dynamic modifications were allowed for each peptide with maximum two repetitions each. Search for phospho-S,T,Y was included only for the IMAC data. Peptide confidence was estimated with the Percolator node. Peptide FDR was set at 0.01 and validation was based on q-value and decoy database search. All spectra were searched against a UniProt fasta file containing 20,165 reviewed human entries. The Reporter Ion Quantifier node included a custom TMT-10plex Quantification Method with integration window tolerance 15 ppm, integration method the Most Confident Centroid at the MS3 level and missing channels were replaced by minimum intensity. Only peptides uniquely belonging to protein groups were used for quantification. Peptide Log$_2$-ratios were computed against the SW48 cell line in each set and were averaged per protein and phosphopeptide. Proteins and phosphopeptides quantified in less than half of the samples were discarded and batch effects were regressed out.
APPENDIX C

SUPPLEMENTARY FIGURES

C.1 Additional results for Chapter 2
C.1.1 Comparison of untargeted metabolomics normalisation to 0 seconds and 25 seconds

Figure C.1: Comparison of untargeted metabolomics normalisation to 0 seconds and 25 seconds. A) Untargeted metabolites fold-changes compared to 0 seconds. B) Metabolites fold-changes compared to 25 seconds. Columns colours denote the different conditions.
C.1.2 Time profiles of representative ions of the untargeted metabolomics with a fold-change compared to time 0 seconds and 25 seconds

**Figure C.2:** Time profiles of representative ions of the untargeted metabolomics with a fold-change compared to time 0 seconds and 25 seconds. A) Time profiles normalised to time zero seconds. B) Time profiles normalised to 25 seconds.
C.1.3 Overlap of significant phosphorylation changes among different conditions

**Figure C.3:** Overlap among the different conditions when considering significantly differentially phosphorylated phosphosites (A, B) and proteins (C, D). A phosphosite is considered to be significantly differentially phosphorylated if it has an adjusted $p$-value lower than 0.05 and an absolute fold-change higher than 2 in at least one time-point. Proteins are considered to be significantly changing in phosphorylation if they have at least one significantly changing phosphosite.
C.1.4 Metabolite connectivity association with measured phosphosites

Figure C.4: Correlation between the number of associated metabolic reactions and the number of significantly differentially phosphorylated site for each of the significantly changing metabolite.

C.2 Additional results for Chapter 3
C.2.1 Agreement between GSEA and Linear regression to estimate enzymatic activities

Figure C.5: Correlation analysis of kinases and phosphatases activities scores between the linear regression and GSEA approaches. A) Correlation of all the kinases/phosphatases activities in each condition. B) Best correlating condition. C) Worst correlating condition.
C.2.2 Principal component analysis of of data-sets and correlation with relative growth rate

Figure C.6: Principal component analysis of data-sets and correlation with relative growth rate. The three principal component with higher absolute correlation coefficient was picked and plotted.
C.2.3 Top predicted metabolites

Figure C.7: List of top predicted metabolites using A) TFs activities and B) kinases/phosphatases activities. List of metabolites that displayed a positive coefficient of determination and significant Pearson correlation between the measured and predicted fold-changes across the different conditions.
C.2.4 ROC-curve analysis of the average feature coefficients

**Figure C.8:** ROC-curve analysis of the average feature coefficients. True-positive tables were built considering the specified resources.

C.3 Additional results for Chapter 4
C.3.1 Replicates distributions of the consumption/release metabolomics experiment

Figure C.9: Distributions of the technical replicates for each biological replicate of the consumption/release experiment.
C.3.2 Hierarchical clustering of the replicates of the consumption/release experiment

Figure C.10: Heatmap of the pairwise pearson correlation coefficient with unsupervised hierarchical clustering of all the technical and biological replicates. Red label represents the samples of the same biological replicate, dark blue represents the WT and light blue the KO cell lines.
C.3.3 Correlation between proteomics and transcriptomics

Figure C.11: Spearman correlation between the proteomics and transcriptomics (RNA-seq) log2 fold-changes (KO - WT).
C.3.4 Unsupervised clustering of the proteomics biological replicates

Figure C.12: Unsupervised clustering of the proteomics biological replicates using all the pairwise pearson correlations.
C.3.5 Hierarchical clustering of the phosphoproteomics replicates

Figure C.13: Heatmap with hierarchical clustering of the pairwise Pearson correlation coefficient of all the phosphoproteomics biological and technical replicates.
C.3.6 Kinases/phosphatases-substrates associations histogram

Figure C.14: Histograms of the number of associations between kinases/phosphatases and substrates. A) Histogram of the number of target phosphorylation-sites of the kinases/phosphatases. B) Histogram of the number of regulatory kinases/phosphatases of the phosphorylation-sites.

C.4 Additional results for Chapter 5
C.4.1 Proteomics and phosphoproteomics batch effect

Figure C.15: Proteomics and phosphoproteomics comparison of the cell lines clustering before and after correcting by batch effect. A) Proteomics cell lines correlation clustering before batch correction and B) after correction. C) and D) similar analysis as in A) and B) using the phosphoproteomics data-set.
C.4.2 Spearman correlation distributions between phosphosites and their protein abundance

**Figure C.16:** Spearman correlation distributions between phosphosites and their protein fold-changes. Dark blue represents the observed correlations and light blue represents the correlations obtained by randomising the matrices maintaining the number and position of missing values.
C.4.3 Unsupervised hierarchical clustering of phosphoproteomics measurements not corrected by protein abundance

Figure C.17: Phosphoproteomics measurements not corrected by protein abundance were used to calculate cell lines pairwise spearman correlations. Unsupervised hierarchical clustering using correlation as a distance metric.
C.4.4 HNF4A association with the proteome

Figure C.18: HNF4A association with the proteome. A) PCA analysis of the proteomics measurements and correlation between the HNF4A protein abundance with principal component 1. B) Significant correlations between transcription-factors and principal component 1.
C.4.5 Proteomics pairwise spearman correlations

Figure C.19: Proteomics pairwise spearman correlations. A) $P$-values distributions of all the unique protein-pairs, protein pairs reported in STRING and protein pairs involved in protein complexes reported in CORUM database. B) Similar to A but distributions represent the correlation coefficients.
C.4.6 Drug response associations with phosphoproteomics and kinases and phosphatases activity

Figure C.20: Drug response associations with phosphoproteomics and kinases and phosphatases activity. A) and C) are QQ-plots of the associations p-values. B) Representative top associations between the phosphosites and drug response. D) Representative top associations between kinases/phosphatases activities and drug response AUC. Lines represent the linear regression between the protein fold-changes and the drug response AUC.


