U.PORTO

DOUTORAMENTO

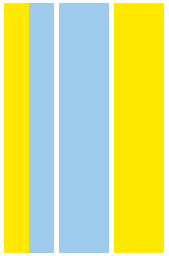
BIOLOGIA BÁSICA E APLICADA

Multi-Omics Characterization of Pan-Cancer Heterogeneity in Gene Expression, Protein Abundance and Protein Activities

Abel Sousa

D

2021



U.PORTO

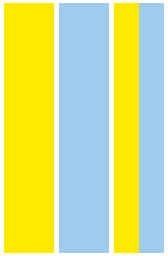
Abel Sousa. Multi-Omics Characterization of Pan-Cancer Heterogeneity in Gene Expression, Protein Abundance and Protein Activities

D.ICBAS 2021

Multi-Omics Characterization of Pan-Cancer Heterogeneity in Gene Expression, Protein Abundance and Protein Activities

Abel Ernesto Fernandes De Sousa

INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
FACULDADE DE CIÊNCIAS
FACULDADE DE MEDICINA



Abel Ernesto Fernandes De Sousa

Multi-Omics Characterization of Pan-Cancer Heterogeneity in Gene Expression, Protein Abundance and Protein Activities

Tese de Candidatura ao Grau de Doutor em Biologia Básica e Aplicada submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Orientador: Professor Doutor Pedro G. Ferreira
Categoria: Professor Auxiliar; Investigador
Afiliação: Faculdade de Ciências da Universidade do Porto (FCUP) - Departamento de Ciência de Computadores; Instituto de Investigação e Inovação em Saúde (i3S); Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC)

Co-orientador: Doutor Pedro Beltrão
Categoria: Investigador Principal
Afiliação: European Bioinformatics Institute (EMBL-EBI)

Co-orientadora: Professora Doutora Carla Oliveira
Categoria: Investigadora Principal; Professora Afiliada
Afiliação: Instituto de Investigação e Inovação em Saúde (i3S); Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Ipatimup); Faculdade de Medicina da Universidade do Porto (FMUP) - Departamento de Patologia

Esta tese foi financiada pela Fundação para a Ciência e a Tecnologia (FCT) através de uma bolsa de doutoramento com a referência PD/BD/128007/2016

Agradecimentos

Uma tese doutoral está longe de ser o fruto de um trabalho individual. Durante o meu percurso académico fui apoiado por muitas pessoas, sem as quais esta tese não teria sido possível.

Em primeiro lugar, quero agradecer aos meus orientadores Pedro Ferreira, Pedro Beltrão e Carla Oliveira. A vocês devo o meu mais sincero obrigado pela confiança que depositaram em mim, dando-me a oportunidade de fazer convosco este doutoramento. Ao longo dos últimos anos mostraram sempre disponibilidade para conversar comigo e discutir ciência. Muito obrigado pelos ensinamentos científicos e por me ajudarem a escolher o melhor caminho a seguir em cada passo. Mais do que isso, obrigado por aquelas palavras de encorajamento que tanto me ajudaram a levar a cabo os meus projetos, e, claro, a escrita desta tese.

A big thank you to all the current and past members of the research groups I was part of: the Expression Regulation in Cancer at i3S and the Cellular Consequences of Genetic Variation at EMBL-EBI. I would like to thank especially to Emanuel, David Ochoa, Danish and Inigo. Many thanks for your guidance and all the answers to whatever questions I had! Inigo thank you a lot for the time you spent teaching me proteomics!

I also need to thank to all the amazing collaborators I met during my projects. Bogdan, Oliver, Aurelien, Borgthor and Julio, your input and contribution was crucial to carry out my projects!

I also want to give a big thank you to all the good friends I made inside and outside the lab during this period of my life. Claudia, Poorya, David Bradley, Inigo, Reham, Christopher and Jorge. Thank you a lot for all your encouraging words and funny moments we have spent together! I am sure I am missing so many people… Please do feel yourselves included!

À Diana e à Helena, a minha principal companhia Portuguesa no EMBL-EBI e em Cambridge, obrigado por me fazerem sentir em “casa”.

A todas as agências de financiamento Portuguesas e Europeias, e, em particular, à Fundação para a Ciência e Tecnologia, agradeço o financiamento que me foi concedido, sem o qual esta tese não teria sido possível.

Quero também agradecer a todas as pessoas envolvidas no meu programa doutoral, o GABBA, e, em particular, aos meus colegas da vigésima edição!

Um grande obrigado aos meus amigos desde os tempos de licenciatura. Como não posso nomear todos, quero destacar o Paulo, o Campos e o Viriato. Obrigado pelas vossas palavras de incentivo e pela energia positiva que sempre me deram e dão. Não me podia esquecer dos meus amigos de infância e adolescência: Hugo, Vítor, Zé, Zito, Paul, Marco, João e Nuno. Vocês foram e serão sempre o meu refúgio na minha vila de Amares. Ao Zé, que, infelizmente, partiu cedo demais, quero agradecer por todas as memórias felizes que me deixou. Numa das últimas vezes que falamos ele disse-me para eu não me preocupar porque a tese iria ficar feita. E ficou. Hoje resta-me apenas sorrir e agradecer-lhe pelo bom amigo que ele foi.

Quero agora agradecer em particular à minha namorada e melhor amiga Susaninha, que foi sem sombra de dúvidas o meu maior pilar durante estes últimos anos. Sem a tua ajuda e as tuas palavras de incentivo não sei se teria chegado ao fim desta aventura. Obrigado por todas as ajudas e discussões científicas que tivemos. Obrigado por teres estado sempre lá para ouvir as minhas lamentações. Obrigado por me teres dado tanta força. Sei que sou um felizardo em ter-te como minha namorada.

Por fim, um agradecimento especial aos meus familiares. Aos meus Pais, ao meu Padrinho, à minha Irmã e à Inês, a minha afilhada e membro mais recente da família, obrigado por me terem apoiado sempre na minha decisão de ir estudar para fora de Portugal. Um agradecimento final aos meus dois cãezinhos Ruca e Kiko, que tornam todo e qualquer momento mais especial e divertido.

Obrigado!

Table of Contents

Abbreviation list

ANOVA	Analysis of variance
ANNOVAR	Annotate variation
AR	Androgen receptor
ATP	Adenosine triphosphate
AUC	Area under the curve
BAM	Binary alignment map
BH	Benjamini-Hochberg
BioGRID	Biological general repository for interaction datasets
Bp	Base pairs
BQSR	Base quality score recalibration
BWA	Burrows-Wheeler alignment
CCL	Cancer cell line
CCLE	Cancer cell line encyclopedia
CDAP	Common data analysis pipeline
cDNA	Complementary DNA
CNV	Copy-number variation
CORUM	The comprehensive resource of mammalian protein complexes
COSMIC	Catalogue of somatic mutations in cancer
CPM	Counts per million
CPTAC	Clinical proteomic tumour analysis consortium
CRC65	65 colorectal cancer cell lines
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPR-cas9	CRISPR-associated protein 9
DE	Differentially expressed
DEG	Differentially expressed gene
DGE	Differential gene expression
DNA	Deoxyribonucleic acid
edgeR	Empirical analysis of digital gene expression data in R
ENCODE	Encyclopedia of DNA elements
eQTL	Expression quantitative trait loci
ER	Estrogen receptor
ES	Enrichment score
FDR	False discovery rate

FPKM	Fragments per kilobase of exon per million reads mapped
GATK	Genome analysis toolkit
GC	Gastric cancer
GDC	Genomic data commons
GEO	Gene expression omnibus
GISTIC	Genomic identification of significant targets in cancer
GLM	Generalized linear model
GO	Gene ontology
GRCh37	Genome reference consortium human genome build 37
GSEA	Gene set enrichment analysis
GTEx	Genotype-Tissue expression project
GWAS	Genome wide association study
HGP	Human genome project
HGVS	Human genome variation society
HPLC	High-performance liquid chromatography
HPM	Human proteome map
HR	Hazard ratio
iBAQ	Intensity based absolute quantification
ICGC	International cancer genome consortium
iPSCs	Induced pluripotent stem cells
IQR	Interquartile range
iTRAQ	Isobaric tag for relative and absolute quantification
JI	Jaccard index
KEGG	Kyoto encyclopedia of genes and genomes
KM	Kaplan-Meier
KSEA	Kinase set enrichment analysis
LC-MS/MS	Liquid chromatography with tandem mass spectrometry
LD	Linkage disequilibrium
Limma	Linear models for microarray data
LoF	Loss-of-function mutation
log10	Logarithm with base 10
log2	Logarithm with base 2
MAF	Mutation annotation format
mRNA	Messenger RNA
MS	Mass spectrometry
MSigDB	Molecular signatures database
NCI60	60 National Cancer Institute human cancer cell lines

NGS	Next-generation sequencing
NHGRI	National human genome research institute
NODE	National omics data encyclopedia
ORA	Over representation analysis
OS	Overall survival
PCA	Principal component analysis
PCAWG	Pan-Cancer analysis of whole genomes consortium
PCR	Polymerase chain reaction
pre-mRNA	Precursor mRNA
PSM	Peptide-spectrum match
PTM	Post-translational modification
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
RPKM	Reads per kilobase of exon per million reads mapped
RPPA	Reverse-phase protein array
rRNA	Ribosomal RNA
RSEM	RNA-Seq by expectation maximization
S	Serine
SAM	Sequence Alignment/Map
SAMseq	Significance analysis of sequencing data
SBG	Sex-biased gene
SD	Standard deviation
Ser	Serine
SILAC	Stable isotope labelling by amino acids in cell culture
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
STAR	Spliced transcripts alignment to a reference
STRING	Search tool for the retrieval of interacting genes/proteins
T	Threonine
TC	Thyroid cancer
TCGA	The cancer genome atlas
TCPA	The cancer proteome atlas
TF	Transcription factor
Thr	Threonine
TMM	Trimmed-mean of M values
TMT	Tandem mass tag

TOM	Topological overlap matrix
TPM	Transcripts per million
tRNA	Transfer RNA
TSG	Tumor suppressor gene
Tyr	Tyrosine
UMAP	Uniform manifold approximation and projection
UV	Ultraviolet
VCF	Variant call format
VEP	Variant effect predictor
VIPER	Virtual inference of protein-activity by enriched regulon analysis
Voom	Variance modeling at the observational level
WGCNA	Weighted gene co-expression network analysis
WES	Whole exome sequencing
WGS	Whole genome sequencing
Y	Tyrosine

Publications

This thesis is supported by the following first author publications:

1. ***Abel Sousa**, Marta Ferreira, Carla Oliveira and Pedro G. Ferreira. **Gender Differential Transcriptome in Gastric and Thyroid Cancers**. Frontiers in Genetics, Volume 11, Article 808, 30 July 2020.*

2. ***Abel Sousa**, Emanuel Gonçalves, Bogdan Mirauta, David Ochoa, Oliver Stegle and Pedro Beltrão. **Multi-omics Characterization of Interaction-mediated Control of Human Protein Abundance levels**. Molecular & Cellular Proteomics, Volume 18, Issue 8, Pages 114-125, 9 August 2019.*

3. ***Abel Sousa**, Aurelien Dugourd, Danish Memon, Borgthor Petursson, Evangelia Petsalaki, Julio Saez-Rodriguez and Pedro Beltrão. **Pan-Cancer Landscape of Protein Activities Identifies Drivers of Signalling Dysregulation and Patient Survival**. BioRxiv, 9 June 2021.*

In addition to the main body of work described in this thesis, I took part in other projects that led to the following publications:

1. *Enio Gjerga, Aurelien Dugourd, Luis Tobalina, **Abel Sousa** and Julio Saez-Rodriguez. **PHONEMeS: Efficient Modeling of Signaling Networks Derived from Large-Scale Mass Spectrometry Data.** Journal of Proteome Research, Volume 20, Issue 4, Pages 2138-2144, 8 March 2021.*

2. *Aurelien Dugourd, Christoph Kuppe, Marco Sciacovelli, Enio Gjerga, Attila Gabor, Kristina B. Emdal, Vitor Vieira, Dorte B. Bekker-Jensen, Jennifer Kranz, Eric.M.J. Bindels, Ana S.H. Costa, **Abel Sousa**, Pedro Beltrao, Miguel Rocha, Jesper V. Olsen, Christian Frezza, Rafael Kramann and Julio Saez-Rodriguez. **Causal Integration of Multi-omics Data With Prior Knowledge to Generate Mechanistic Hypotheses.** Molecular Systems Biology, Volume 17, Issue 1, 1 January 2021.*

3. *Marta Moreno, **Abel Sousa**, Marta Melé, Rui Oliveira and Pedro G Ferreira. **Predicting Gastric Cancer Molecular Subtypes from Gene Expression Data**. Proceedings, Volume 54, Issue 1, 7 September 2020.*

4. *Andrea Rebeca Bustos-Carpinteyro, Carla Oliveira,* ***Abel Sousa****, Patricia Oliveira, Hugo Pinheiro, Joana Carvalho, María Teresa Magaña-Torres, María Guadalupe Flores-Miramontes, Adriana Aguilar-Lemarroy, Luis Felipe Jave-Suárez, Jorge Peregrina-Sandoval, José Alfonso Cruz-Ramos & Josefina Yoaly Sánchez-López.* ***CDH1 somatic alterations in Mexican patients with diffuse and mixed sporadic gastric cancer.*** *BMC Cancer, Volume 19, Issue 1, Pages 1-9, 14 January 2019.*

5. *Marta Dueñas, Andrés Pérez-Figueroa, Carla Oliveira, Cristian Suárez-Cabrera,* ***Abel Sousa****, Patricia Oliveira, Felipe Villacampa, Jesús M Paramio, Mónica Martínez-Fernández.* ***Gene Expression Analyses in Non Muscle Invasive Bladder Cancer Reveals a Role for Alternative Splicing and Tp53 Status.*** *Scientific Reports, Volume 9, Issue 1, Pages 1-11, 17 July 2019.*

6. *Pedro G Ferreira, Manuel Muñoz-Aguirre, Ferran Reverter, Caio P Sá Godinho,* ***Abel Sousa****, Alicia Amadoz, Reza Sodaei, Marta R Hidalgo, Dmitri Pervouchine, Jose Carbonell-Caballero, Ramil Nurtdinov, Alessandra Breschi, Raziel Amador, Patrícia Oliveira, Cankut Çubuk, Joao Curado, François Aguet, Carla Oliveira, Joaquin Dopazo, Michael Sammeth, Kristin G Ardlie, Roderic Guigó.* ***The effects of death and post-mortem cold ischemia on human tissue transcriptomes.*** *Nature Communications, Volume 9, Issue 1, Pages 1-15, 13 February 2018.*

Abstract

Cancer is a highly heterogeneous disease that can be caused by the acquisition of somatic DNA mutations, including single nucleotide variants, gene copy-number variations (CNVs) and large chromosomal rearrangements. The Cancer Genome Atlas (TCGA) has led to an in-depth characterization of the genomic alterations of more than 10,000 tumours from 33 cancer types. These mutations generate the genetic diversity that promotes the acquisition of multiple cancer hallmarks, including chronic proliferation, resistance to cell death and tissue invasion and metastasis. Due to technical limitations, the study of protein abundances and signalling activities has been for many years limited primarily to the study of a few key signalling proteins at a time via the use of reverse-phase protein arrays (RPPA). The Clinical Proteomic Tumour Analysis Consortium (CPTAC) has revolutionized the study of cancer proteomes, including proteins and respective post-translational modifications, through the application of Mass Spectrometry-based proteomics.

An understanding of the molecular mechanisms that underpin the development of cancer is critical in order to study cancer biology and to develop therapies. However, the study of the cellular consequences of cancer genetic alterations was for many years limited to few key cancer driver genes. TCGA and CPTAC have revolutionized the study of the molecular basis of cancer. It is now possible to perform systematic Pan-Cancer characterizations of the impact of genomic variation on multiple layers of biological information. Moreover, these projects allow cancer researchers to study other sources of molecular heterogeneity in cancer, including the impact of patient phenotypes such as gender. The main goal of this thesis is to leverage multi-omics cancer datasets from TCGA and CPTAC in order to build an integrated picture of the impact of genomic alterations and gender on gene expression, protein abundance and protein activities across multiple cancer types.

Cancer has an important and considerable gender differential susceptibility confirmed by several epidemiological studies. Beyond environmental predisposing factors, gender intrinsic molecular characteristics may also play an important role. In this thesis, we performed a deep gender-differential gene expression and co-expression network analysis in malignant and non-malignant tissues of stomach and thyroid, two tissue types with unbalanced cancer incidences between genders. We found that sex-biased gene expression is more pronounced in normal tissues than tumour tissues and that most of the shared variation arises from the sexual chromosomes. Expression of several cancer-associated genes differs between genders, with tumour suppressor genes preferentially downregulated in the tumour tissue of the most susceptible gender. Gene co-expression

network analysis revealed an extensive topological preservation between genders, with gender-specific networks appearing correlated with cancer histological subtypes.

Gene copy-number changes are widespread across many forms of human cancer and often act as driver events. However, the effects of CNVs on the proteome of tumours are poorly understood. To study the propagation of CNVs to the mRNA and protein levels of cancer cells, we performed a multi-omics analysis that combines genomics, transcriptomics, (phospho)proteomics, and structural data. Analysing 8,124 proteins, we observed that up to 42% of them show evidence of protein-level gene dosage attenuation. Over 500 protein-protein interactions show indirect control of degradation of one subunit via physical associations, 32 of which may be further controlled by phosphorylation. Using structural models for 3,082 protein interfaces, we found that a higher fraction of interface residues is associated with a higher degree of attenuation. Finally, we studied the impact of these findings on non-malignant cell samples. We found evidence of interaction-mediated control of protein abundances in normal tissues. Moreover, the degree of protein attenuation correlates with the probability that natural genetic variation with an impact on gene expression may result in a phenotypic consequence.

Somatic DNA mutations in cancer cells are strongly linked to alterations in kinase and transcription factor (TF) activities. While the mutational profiles of diverse tumours have been extensively characterized, until recently, the measurements of protein activities have been technically limited to a few proteins at a time. To perform a systematic investigation of the impact of genomic alterations on the activities of kinases and TFs, we mined multi-omics datasets made available by the TCGA and CPTAC consortia and cancer cell line studies. In this study, we estimated changes in the activities of 218 kinases and 292 TFs from gene expression and phosphoproteomics data of 1,110 tumours and 77 cell lines. Predicted kinase activities were supported by their agreement with functionally annotated phosphosites and RPPA phosphorylation data. Co-regulation of kinase and TF activities reflects previously known regulatory relationships and allows dissecting genetic drivers of signalling changes in cancer. Finally, we identified the activities most often regulated in cancer and showed how these can be linked to differential patient survival. Altogether, the protein activity profiles across over 1000 cancer samples serve as a resource to study the dysregulation of signalling across different tumour types.

Resumo

O cancro é uma doença altamente heterogénea que pode ser causada pela aquisição de mutações somáticas no DNA, incluindo variantes de nucleótido único, variações no número de cópias dos genes (CNVs) e extensos rearranjos cromossómicos. O Atlas do Genoma do Cancro (TCGA) levou a uma caracterização profunda das alterações genómicas de mais de 10,000 tumores de 33 tipos diferentes de cancro. Estas mutações geram a diversidade genética necessária para a obtenção de múltiplas características carcinogénicas, como por exemplo proliferação constitutiva, resistência à morte celular programada e invasão de tecidos adjacentes e metastização. Devido a limitações técnicas, os estudos das abundâncias proteicas e atividades de sinalização foram por muitos anos limitados a algumas proteínas consideradas chave nos processos de sinalização celular. Esta limitação deveu-se principalmente ao uso de *microarrays* de proteínas em fase reversa (RPPA). O Consórcio Clínico de Análise Proteica de Tumores (CPTAC) revolucionou o estudo dos proteomas de cancro, incluindo proteínas e respetivas modificações pós-tradução, através do uso de técnicas de proteómica baseadas em espectrometria de massa.

A caracterização dos mecanismos moleculares que estão na base do desenvolvimento do cancro é crucial para estudar a biologia do cancro e desenvolver terapias mais sofisticadas. Contudo, o estudo das consequências celulares das alterações genéticas foi por muitos anos limitado a alguns genes que são considerados chave para o desenvolvimento e progressão do cancro. Os consórcios TCGA e CPTAC revolucionaram o estudo da base molecular do cancro. Hoje é possível efetuar caracterizações sistemáticas do impacto da variação genómica em múltiplos tipos de moléculas e ao longo de vários tipos de cancro. Além disso, estes consórcios permitem que investigadores na área do cancro investiguem outras fontes de heterogeneidade molecular, incluindo o sexo dos pacientes. O principal objetivo desta tese é usar conjuntos de dados *multi-omics* para obter uma visão integrada do impacto de alterações genéticas e do sexo dos pacientes na expressão dos genes, na abundância proteica e na atividade de proteínas ao longo de vários tipos de cancro.

O cancro é uma doença que tem uma suscetibilidade diferencial elevada entre sexos que foi confirmada por vários estudos epidemiológicos. Para além de fatores ambientais de predisposição, as características moleculares intrínsecas a cada sexo também podem contribuir para esta incidência distinta. Nesta tese, nós fizemos uma análise detalhada das diferenças de expressão dos genes e redes de co-expressão entre géneros em tecidos malignos e não malignos de estômago e tiroide. Estômago e tiroide

foram escolhidos devido à carcinogénese diferencial elevada entre géneros. Neste estudo, nós verificámos que os tecidos normais apresentam mais diferenças de expressão génica do que os tecidos tumorais e que a maioria da variação partilhada é proveniente dos cromossomas sexuais. Para além disso, nós descobrimos que a expressão de vários genes associados ao cancro difere entre géneros, e que genes supressores tumorais estão preferencialmente sub-regulados no tecido tumoral do género mais suscetível ao cancro. A análise de redes de co-expressão revelou uma extensa preservação topológica entre géneros. Não obstante, foram encontradas redes específicas de cada género correlacionadas com os subtipos histológicos de cancro.

As alterações no número de cópias dos genes são frequentes em vários tipos de cancro e estão associadas diretamente com a tumorigénese. Todavia, o efeito dos CNVs no proteoma dos tumores ainda está pouco descrito. De modo a estudar a propagação dos CNVs nos níveis de proteína e mRNA dos tumores, nós fizemos uma análise *multi-omics* que combina genómica, transcritómica, (fosfo)proteómica e dados estruturais de proteínas. A partir da análise de 8,124 proteínas, nós descobrimos que cerca de 42% das proteínas têm evidências de atenuação dos efeitos do número de cópias dos genes. Para além disso, nós encontramos mais de 500 interações proteína-proteína que mostram que uma subunidade proteica tem a capacidade de controlar indiretamente a degradação de outra subunidade através de interações físicas. Destas mais de 500 interações, 32 podem ainda ser controladas por diferenças de fosforilação. A partir de 3,082 modelos estruturais de interfaces proteicas, nós constatámos que as proteínas que têm uma maior atenuação do número de cópias dos genes apresentam uma fração maior de resíduos em interfaces. Por último, nós estudámos o impacto destas descobertas ao nível de amostras celulares não malignas. Estas análises permitiram encontrar evidências em tecidos normais de controlo de abundância proteica mediado por interações físicas. Além disso, o grau de atenuação está correlacionado com a probabilidade de que variação genética natural com impacto na expressão dos genes possa ter consequências fenotípicas.

As mutações somáticas no DNA das células cancerígenas estão fortemente associadas com alterações nas atividades das quinases e fatores de transcrição (TFs). Enquanto que os perfis mutacionais de diversos tumores têm sido amplamente caracterizados, a quantificação de atividades proteicas tem sido limitada a um número reduzido de proteínas devido a razões técnicas. De modo a investigar sistematicamente o impacto de alterações genómicas nas atividades de quinases e fatores de transcrição, nós compilámos conjuntos de dados *multi-omics* partilhados pelo TCGA e CPTAC e estudos em linhas celulares de cancro. Neste estudo, nós estimámos as atividades de 218 quinases e 292 fatores de transcrição a partir de dados de expressão dos genes e fosforilação de proteínas de 1,110 tumores primários e 77 linhas celulares. As atividades de quinases

previstas foram suportadas pela concordância com locais de fosforilação anotados funcionalmente e dados de fosforilação de RPPA. Para além disso, a co-regulação entre quinases e TFs foi refletida em relações regulatórias previamente conhecidas e permitiu explorar alterações genéticas que causam variações de sinalização em cancro. Por último, nós identificámos as atividades proteicas que estão mais frequentemente reguladas em cancro e como estas estão associadas com a sobrevivência dos pacientes. Em suma, os perfis de atividade proteica ao longo de mais de 1,000 amostras de cancro constituem um recurso para estudar a desregulação da sinalização celular em tumores.

1. Introduction

1.1. Multi-omics approach to molecular biology

At the time of writing of this thesis, the suffix *omics* is largely used in biology to describe the comprehensive and high throughput quantification of functionally-related biochemical molecules. These molecules have functions that define the structure, homeostasis and dynamics of the organisms (Hasin, Seldin, and Lusis 2017). Beyond the high-throughput quantification of chemical species or the technologies involved in these processes, omics also refers to independent fields of research (Medicine 2012). The closely related suffix *ome* is in turn used to refer to the object of study of such fields, e.g., genomics refers to the study of the genome - the genetic makeup of an organism. In fact, since Dr. Thomas H. Roderick, a geneticist at The Jackson Laboratory, used the term genomics for the first time in 1986, there has been an explosion of omics fields (Yadav 2007). The most well-known omics fields are: genomics, epigenomics, transcriptomics, proteomics, phosphoproteomics and metabolomics, which correspond to the study of the genes and genomes, epigenetic modifications, RNA, proteins, phosphorylation of proteins and metabolites, respectively (Medicine 2012). These research fields allowed for great advances in the computational modelling of complex biological systems, an interdisciplinary field of research usually known as *systems biology* (**Figure 1.1**). As such, novel bioinformatics tools have been emerging in the past years in order to circumvent the challenges of multi-omics data integration and interpretation (Argelaguet et al. 2018, 2020; Dugourd et al. 2021).

Omics technologies allowed researchers to explore the molecular basis of many diseases at unprecedented resolution (Hasin, Seldin, and Lusis 2017). Cancer research is one of the areas that largely benefited from such advances. High-throughput technologies fostered the development of major cancer genomic and proteomic projects that exponentially improved our understanding of how cancer cells behave. Moreover, these projects led to the generation of an exciting amount of multidimensional datasets that simultaneously profile multiple molecular layers of biological information (e.g., DNA, RNA and proteins). Importantly, this data allows other cancer researchers to computationally address *a posteriori* a myriad of cancer-related questions. This thesis aims to leverage the power of multi-omics datasets to study the impact of genomic alterations and patient phenotypes on the gene expression, protein abundance and protein activities of cancer cells.

The introductory chapter of this thesis is organized as follows: in the next subchapter (**1.2**) I discuss the cancer genomic and proteomic consortia that revolutionized the study of cancer. In subchapter (**1.3**), I describe sources of heterogeneity in gene expression, protein

abundance and protein activities in cancer. After that, I address in detail the omics fields that are more relevant to this thesis: genomics (**1.4**), transcriptomics (**1.5**), proteomics (**1.6**) and phosphoproteomics (**1.7**). In these subchapters, I make a comprehensive overview of the main technological details of data acquisition, as well as methods of data pre-processing and data mining. The specific aims of this thesis are finally enumerated in subchapter **1.8**.

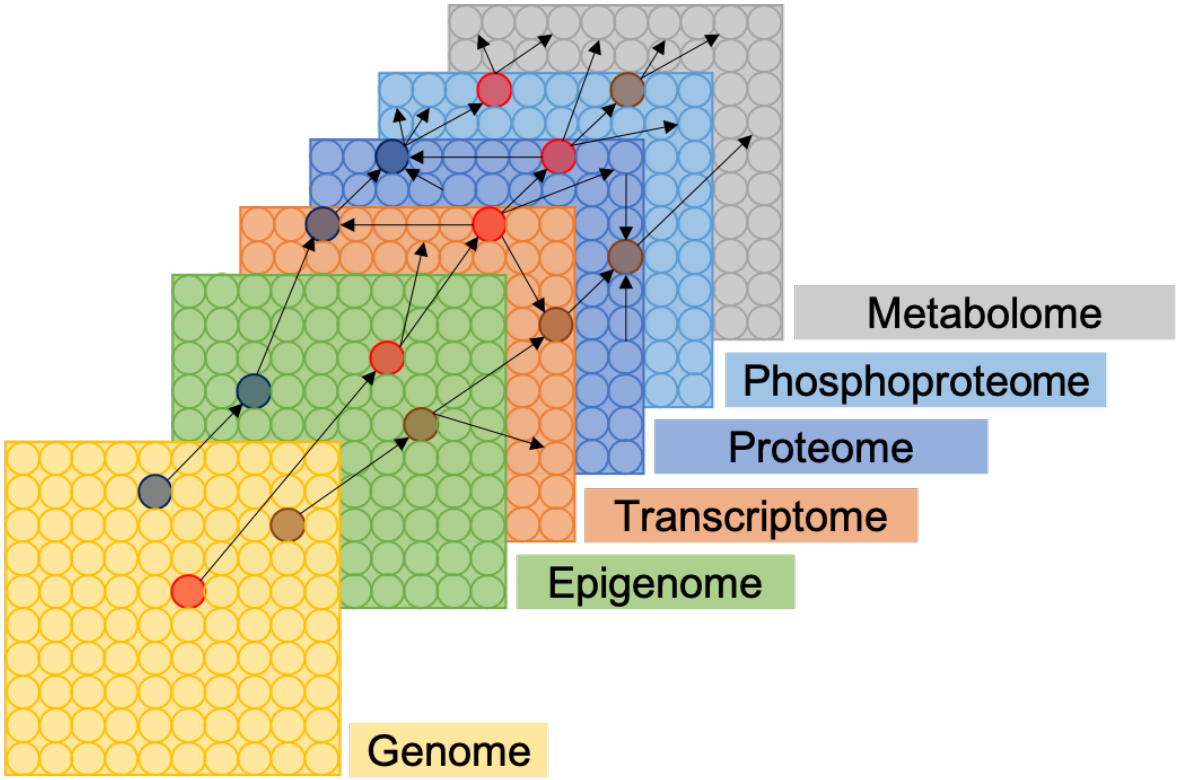


Figure 1.1. Multi-omics datasets and interactions between molecular layers. Each layer represents a category of biological molecules assayed using omics technologies: genome, epigenome, transcriptome, proteome, phosphoproteome and metabolome. The data is collected from the entire pool of molecules (genes, transcripts, proteins, etc.), represented here as circles. Three genes were highlighted for representation purposes: blue, red and brown. The solid black arrows represent potential interactions or correlations between molecules. These interactions occur between molecules from the same layer (e.g., RNA-RNA and protein-protein correlations) or from different layers (e.g., DNA-RNA/protein, DNA methylation-RNA, RNA-protein, protein-phosphosite, protein-metabolite correlations). Systems biology aims to model and disentangle the complex network of intra- and inter-layer interactions. Figure adapted from (Hasin, Seldin, and Lusis 2017).

1.2. Large-Scale cancer genomic and proteomic projects

Cancer has been addressed as a disease of the genome (Vazquez, de la Torre, and Valencia 2012; B. Zhang et al. 2019). There is a large pool of potentially pathogenic somatic

and germline DNA mutations that can alter once well-controlled cellular pathways and make the cells divide abnormally. In the long term, the uncontrolled cellular growth can originate a neoplasm, also known as a tumour. A malignant tumour has the capacity of spreading and evading other tissues, a condition that is usually called cancer. Genome sequencing studies emphasized the importance of identifying and cataloguing tumour-associated mutations. The advances made in the next-generation sequencing (NGS) technologies over the last two decades underpinned the establishment and development of cancer genomic projects, such as the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). Throughout the last decade, TCGA has profiled and analyzed the genome, epigenome, transcriptome and, to some extent, the proteome of over 10,000 tumours from 33 cancer types (Hoadley et al. 2018; Ding et al. 2018). This remarkable effort led to enormous advancements in our understanding of the oncogenic processes governing cancer development and progression. TCGA helped to: (i) improve the stratification of cancer patients based on their molecular profiles. These cancer molecular subtypes complement the classical histopathological classifications and encouraged the development of anti-cancer therapies targeting distinct populations of patients, which can ultimately improve survival (Bass et al. 2014; Koboldt, Fulton, et al. 2012; Bell et al. 2011; Muzny et al. 2012); (ii) elucidate the cross-talk between the germline and the somatic cancer genome. It was found that the germline genome has far-ranging influences on the somatic landscape and often promotes somatic mutations (Ding et al. 2018); (iii) redefine the catalogue of cancer driver genes and characterize the *cis/trans* effects of driver mutations on gene expression (M. H. Bailey et al. 2018); (iv) compare the prevalence of tumorigenic processes, such as the dysregulation of cellular proliferation, death and adhesion, across cancer types (Ding et al. 2018); (v) change the paradigms of cancer research, allowing to study the tumours within the context of their microenvironments rather than in isolation. This has proved to be especially important in the study of tumour-infiltrating immune cells and in the development of new treatments and immunotherapies (Thorsson et al. 2018).

Despite the profound breakthroughs of TCGA in illuminating the genomic landscape of human malignancies, the cancer proteome remained poorly understood. The cancer proteome corresponds to the qualitative and quantitative map of all expressed proteins, protein complexes and post-translational modifications (PTMs) in a tumour (Nesvizhskii 2014). The first large-scale efforts to characterize the protein and PTM profiles of tumours were done by TCGA using reverse-phase protein arrays (RPPA) (Akbani, Ng, et al. 2014). However, this technique requires *a priori* knowledge about the proteins of interest and has a relatively low throughput due to the dependency on the quality and number of antibodies available (Weinstein et al. 2013; Alfaro et al. 2014). Besides the obvious biological difficulties in mapping the complex and highly-dynamical cancer proteome, the progress of

cancer proteomics has also been hampered by the greater technical challenges of measuring proteins rather than DNA or RNA (Paulovich and Whiteaker 2016). Nevertheless, changes in protein regulation should be more closely linked to alterations in cellular processes and phenotypes. Therefore, the integration of the tumour-derived genome, transcriptome and proteome, i.e., proteogenomics, is crucial for a better understanding of cancer biology (Alfaro et al. 2014). With this idea in mind, cancer researchers established the Clinical Proteomic Tumour Analysis Consortium (CPTAC). CPTAC has revolutionized the study of the molecular basis of cancer through the application of Mass Spectrometry (MS)-based proteomics and phosphoproteomics. MS is an unbiased and rapidly evolving high-throughput technique that allows the accurate identification and quantification of tens of thousands of proteins nowadays. Additionally, it allows to map the correct localization of PTMs, such as phosphorylation, acetylation and glycosylation (Edwards et al. 2015). CPTAC has provided a comprehensive proteogenomic characterization of genomically annotated TCGA tumours (B. Zhang et al. 2014; Mertins et al. 2016; H. Zhang et al. 2016) and, more recently, of independent cancer cohorts (Gillette et al. 2020; Dou et al. 2020; Clark et al. 2019). These studies demonstrated the added value of proteomics and phosphoproteomics data for the classical genomics-driven cancer research. CPTAC enabled to: (i) identify additional cancer molecular subtypes not detectable by transcriptomic and genomic profiles; (ii) find that changes at genomic and transcriptomic level are often buffered at the proteomic level, which can help to prioritize candidate cancer driver genes; (iii) elucidate the functional consequences of somatic mutations by proteomic validation; (iv) uncover the landscape of dysregulated signalling pathways by phosphoproteomics data integration; (v) associate the proteomic and phosphoproteomic-derived signatures with patients outcomes, including the overall survival of cancer patients; (vi) discover novel prognostic cancer biomarkers and drug targets (Chen et al. 2019; B. Zhang et al. 2019).

1.3. Cellular consequences of cancer mutations and patient phenotypes

The repertoire of cancer genomic alterations can have profound impacts on the physical and chemical components of the cancer cell. According to the central dogma of molecular biology, the genetic information flows from DNA to RNA, and from RNA to proteins. RNA molecules are therefore the first cellular players to be affected by cancer mutations, resulting in changes in gene expression and in alternative splicing (Calabrese et

al. 2020). However, cancer mutations can have far-reaching cellular consequences and also affect protein homeostasis and signalling transduction pathways through the deregulation of kinase activities (Akbani, Ng, et al. 2014; Blume-Jensen and Hunter 2001).

Previous studies in breast cancer found that expression of *ERBB2* RNA and its protein product HER2 are well correlated with *ERBB2* DNA amplifications (Slamon et al. 1989; Owens, Horten, and Da Silva 2004). In addition, other studies elucidated how HER2 overexpression translates into signals that potentiate dysregulated cell growth and oncogenesis in breast cancer (Harari and Yarden 2000). It was found that HER2 overexpression is associated with high autophosphorylation activity of this receptor tyrosine kinase (Lonardo et al. 1990). Moreover, mutant HER2 strongly interacts with the MAPK (Ben-Levy et al. 1994) and the PI3K pathways (Peles et al. 1992), supporting the conclusion that HER2 stimulates cell proliferation and survival through these pathways. To corroborate this hypothesis, signalling through the MAPK pathway is significantly prolonged and enhanced in breast cancer cells overexpressing HER2 (Karunagaran et al. 1996).

The advent of high-throughput *omics* technologies provided an opportunity to scale these pioneer studies up to the level of the whole genome and across multiple cancer types: it is now possible to perform Pan-Cancer systematic characterizations of the impact of genomic variation on gene expression, protein abundance and protein activities. Moreover, these technologies allow cancer researchers to study other sources of molecular heterogeneity in cancer, including the impact of epigenetic states and patient characteristics such as age and gender. The main goal of this thesis is to exploit multi-omics cancer datasets to build an integrated picture of the impact of genomic alterations, and patient-specific characteristics, on the RNAs and proteins of cancer cells.

The following subchapters provide a mechanistic description of how cancer mutations and patient-specific endogenous factors contribute to variation in gene expression, protein abundance and protein activities. In the next subchapter (**1.3.1**) I give emphasizes to the role of genetic alterations and sex in shaping the cancer transcriptome. After that, I describe the molecular basis of sex differences in cancer incidence and survival. In subchapter **1.3.2** I discuss how genetic changes in the form of copy-number alterations are not always reflected at the protein level, and what are the mechanisms that underlie this buffering effect. Finally, in subchapter **1.3.3** I introduce the effects of mutations on protein activities and cell signalling pathways.

1.3.1. Determinants of gene expression variability in cancer

Gene expression variation in cancer is largely driven by somatic genomic alterations, including gene copy-number variations (CNVs) and single nucleotide variants (SNVs), epigenetic changes and microRNA-mediated post-transcriptional regulation (Sharma, Jiang, and De 2018). In a recent study from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium, it was found that CNVs are the major drivers of gene expression variation (17%), followed by somatic SNVs in proximal gene regions (2%) (Calabrese et al. 2020). Regarding the impact of SNVs in gene expression, multiple studies reported hundreds to thousands of variants in cancer genomes that are correlated with gene expression levels (Ongen et al. 2014; Gong et al. 2018; Geeleher et al. 2018; Calabrese et al. 2020). These variants are termed expression quantitative trait loci (eQTLs) and are classified as either *cis* (local) or *trans* (distal) depending on the location of the associated gene relative to the variant. Some eQTLs have been further associated with patient overall survival (Gong et al. 2018). It has been hypothesized that the diversity of molecular aberrations (e.g., gene expression) across cancer types is the result of endogenous factors, such as tissue-specific developmental and differentiation programs and epigenetic states, in conjunction with exogenous factors, including mutagenic exposures and inflammation. In accordance with this hypothesis, it was found that the cell-of-origin of more than 10,000 tumours was largely recapitulated by mRNA-based unsupervised clustering (Hoadley et al. 2018).

Gender is an example of an endogenous factor and phenotype that can influence gene expression in normal tissues (Aguet et al. 2020; Lopes-Ramos et al. 2020) and in cancer (Yuan et al. 2016; J. Ma, Malladi, and Beck 2016). For this reason, I investigated the sex-biased gene expression patterns in malignant and non-malignat tissues of stomach and thyroid, two tissue types with unbalanced cancer incidences between genders (Siegel, Miller, and Jemal 2018; Rahbari, Zhang, and Kebebew 2010). This study is shown in chapter **2**. The rationale behind sex disparities in cancer is introduced in the following subchapter.

1.3.1.1. The role of sex differences in cancer incidence

It has been hypothesized that sex-specific gene regulation underlies important phenotypic gender differences and may contribute to gender-differential susceptibility to disease (Ober, Loisel, and Gilad 2008; Rawlik, Canela-Xandri, and Tenesa 2016; Labonté et al. 2017). Cancer has a considerable differential incidence between genders (Tevfik Dorak and Karpuzoglu 2012; Clocchiatti et al. 2016; Ali et al. 2016), with men showing

higher cancer incidence than women in 32 of 35 anatomical sites (Edgren et al. 2012). In 13 of these sites, the differences could not be explained by known risk factors, including smoking, alcohol consumption and potential occupational carcinogens such as toxic metals and ionizing radiation. Men are at higher risk and worst prognosis in several types of cancers in non-reproductive tissues, including skin, esophagus, stomach, liver, and urinary bladder cancers (Siegel, Miller, and Jemal 2018). One remarkable exception is the thyroid tissue, where women have three times higher risk of developing cancer (Rahbari, Zhang, and Kebebew 2010). For malignancies such as acute lymphoblastic leukemia (ALL) or non-Hodgkin lymphoma, the gender-biased incidence occurs already in childhood, being more common in boys (Tevfik Dorak and Karpuzoglu 2012). Although environmental and lifestyle factors largely contribute to gender disparities in cancer, it seems clear that gender intrinsic molecular factors may also play an important role. The molecular basis of cancer sexual disparity may be the consequence of a complex interplay between sex chromosomes and the hormonal system (Clocchiatti et al. 2016) (**Figure 1.2**).

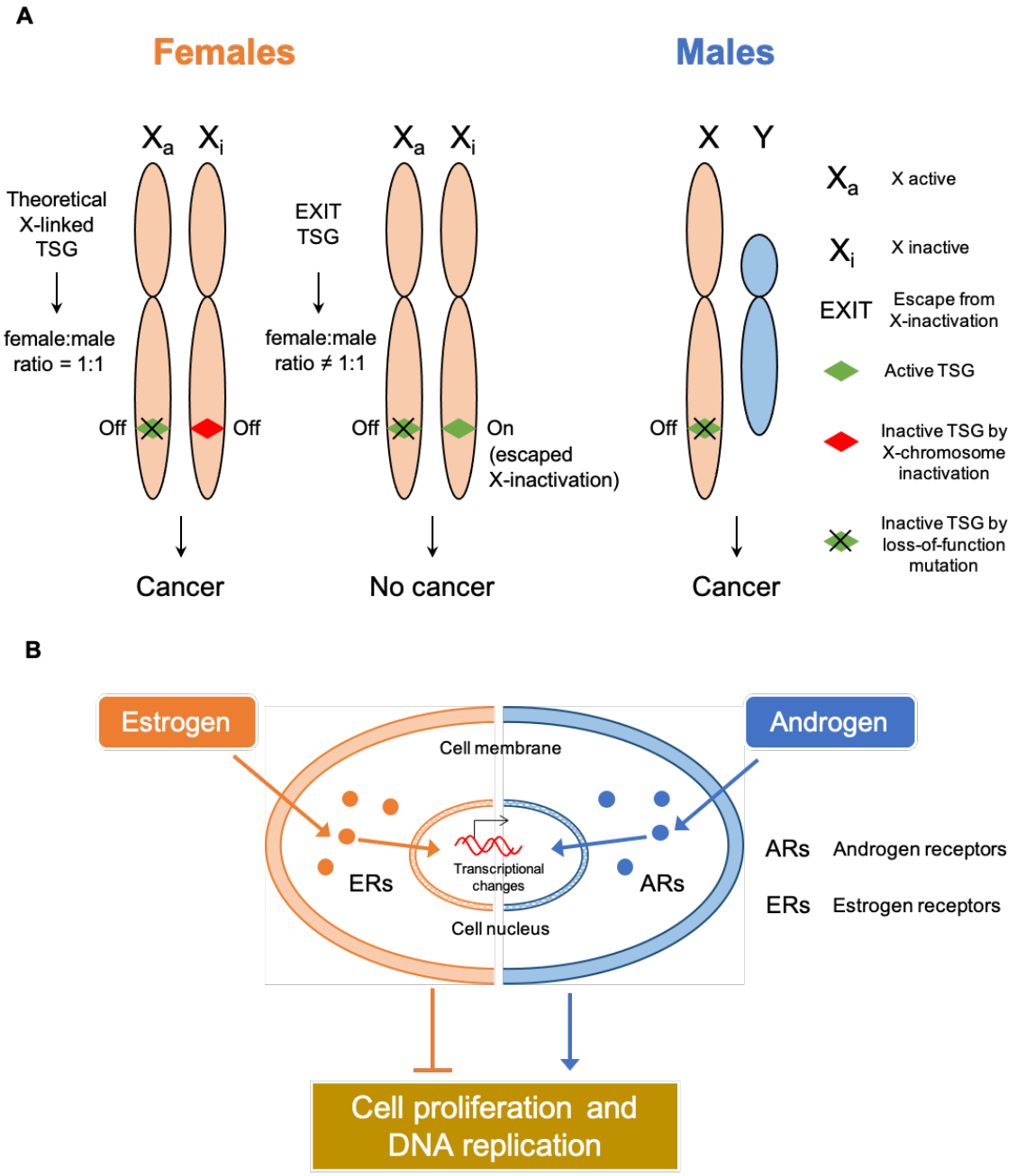


Figure 1.2. Possible mechanisms contributing to gender differential susceptibility to cancer. (A) Females and males differ on their sex chromosomes: females have two X chromosomes (XX) and males have one X chromosome and one Y chromosome (XY). During female embryogenesis, one X chromosome is inactivated (X_i) by a specialized RNA-based silencing mechanism to prevent gene-dosage imbalances between genders. Therefore, only one copy is transcriptionally active (X_a) in females, balancing the expression with males. According to this model, both sexes are equally susceptible to inactivating mutations in X-linked tumour suppressor genes (TSGs): males and females would only require a single deleterious mutation to lose a X-linked TSG, which might in turn contribute to carcinogenesis. However, some genes, including TSGs, escape from X-inactivation (EXIT), protecting females from complete TSG loss after a single loss-of-function mutation. In this model, complete TSG inactivation would require biallelic mutations, or mutation with loss of the other X chromosome. **(B)** Androgens and estrogens flow inside the cell through the cytoplasmic membrane and bind to the corresponding steroid receptors: estrogen receptors (ERs) and androgen receptors (ARs), respectively. Hormone binding to the

receptor triggers alterations in structural conformation and nuclear localization. The receptors then bind to specific steroid-response elements in the DNA to activate or repress the transcription of genes. In some cancer models it was found that androgens and estrogens might have antagonistic effects regarding cancer development and progression. In liver cancer (more common in males), androgen and estrogen signalling have opposite effects on hepatocyte proliferation and DNA replication, where AR stimulates and ER restrains cell division (Z. Li et al. 2012). These hormones have contrasting effects in the thyroid tissue (Stanley et al. 2012; M. L. Lee et al. 2005), which may drive the higher incidence of thyroid cancer in females. Figure adapted from (Dunford et al. 2017; Clocchiatti et al. 2016; Levin and Hammes 2016).

1.3.1.1.1. Sex chromosomes

Females have two X chromosomes, one inherited from each parent, whereas males carry one X chromosome inherited from the mother and one Y chromosome from the father (Libert, Dejager, and Pinheiro 2010). To ensure a correct dosage of X-linked genes between both sexes, one copy of the X chromosome is usually randomly inactivated by the long non-coding RNA *XIST* in the early stages of female embryogenesis. However, some genes may escape the *XIST*-dependent inactivation, triggering an imbalanced expression between males and females (Carrel and Willard 2005; Tukiainen et al. 2017). This asymmetric expression can make females more resistant to inactivating mutations in X-linked tumour suppressor genes (TSGs), as males would require only a single deleterious mutation and females would require two (Dunford et al. 2017) (**Figure 1.2A**).

UTX is a TSG (encoding a histone H3K27 demethylase) that is known to escape silencing in females (Van Haaften et al. 2009; Greenfield et al. 1998; Bellott et al. 2014). Males have a *UTX* paralogous on the Y chromosome called *UTY*, however its catalytic activity is relatively low and may not compensate for inactivating mutations in the *UTX* gene (Walport et al. 2014). This seems to be important in the aetiology of ALL, which has been reported to have somatic loss-of-function mutations in *UTX* (Van Der Meulen et al. 2015; Mar et al. 2012). As described above, ALL is prevalent in childhood and more common in males than females. Additionally, it was reported inactivating *UTX* mutations in renal and esophageal cancers, which are more common in males (Van Haaften et al. 2009; Dalgliesh et al. 2010; Lucca et al. 2015; S. H. Xie and Lagergren 2016). Moreover, Dunford *et al.* recently reported that four TSGs that escape from X-inactivation, including *UTX*, harboured loss-of-function mutations more frequently in male cancers (Dunford et al. 2017). The authors concluded that biallelic expression of X-linked TSGs in females partially explains the reduced cancer incidence in this sex across multiple tissue types.

The X chromosome also contains a high number of genes directly or indirectly involved in immunological activity, and females have better immune responses to pathogens than males (Libert, Dejager, and Pinheiro 2010; Klein and Flanagan 2016). The increased immune activity in females is likely to be accompanied by enhanced cancer immunosurveillance (Clocchiatti et al. 2016), at the expense of an increased susceptibility to autoimmunity (Klein and Flanagan 2016; Mousavi, Mahmoudi, and Ghotloo 2020).

1.3.1.1.2. Sex hormones

Sex steroid hormones enter the cells through the plasma membrane and bind to cellular receptors, such as the estrogen receptor-α (ERα), ERβ and androgen receptor (AR). Next, the steroid-bound receptors translocate to the cell nucleus and regulate the transcription of the target genes, affecting cellular metabolic states, the immune system, cancer stem cell self-renewal and tumour microenvironments (Levin and Hammes 2016; Clocchiatti et al. 2016) (**Figure 1.2B**). Several cancer types are more incident and aggressive in males and postmenopausal females compared with premenopausal females, including cancers of the liver, colon, kidney, oesophagus, skin, and head and neck (Clocchiatti et al. 2016). Higher estrogen signalling, leading to ERβ activation, has been implicated as one of the factors behind female protection in cancer. To corroborate this hypothesis, ERβ levels are often diminished in multiple types of cancers, and sustained ERβ expression is a favourable prognostic marker in renal cancer (C. P. Yu et al. 2013). In colon cancer, a polymorphism in the promoter of *ESR2* (gene encoding ERβ), which likely alters its expression, was associated with increased survival of postmenopausal women (Passarelli et al. 2013).

Hepatocellular carcinoma (HCC) is more common in males in both rodents and humans (Kalra et al. 2008; Sung et al. 2021). In mouse models of the disease, it was found that male mice treated with estrogen developed fewer tumours than control males, and ovariectomized females developed more tumours than normal females during chemically-induced carcinogenesis (Naugler et al. 2007; Shimizu et al. 1998; Tsutsui et al. 1992; Yamamoto et al. 1991). In addition, reduced incidence of HCC was observed in male mice not expressing AR (C. L. Ma et al. 2008; M. H. Wu et al. 2010). These results demonstrated that estrogens and androgens contribute to hepatocarcinogenesis in an antagonistic manner (**Figure 1.2B**). A more detailed analysis suggested this opposite role may happen through a *FOXA1/2*-dependent regulation of gene expression: AR stimulates whereas ERα

restrains cellular proliferation and metabolism of nucleotides and amino acids (Z. Li et al. 2012).

As previously noted, the incidence of thyroid cancer is significantly higher in females than in males. Although the reasons for this disparity are still largely unclear, it was reported that AR expression in thyroid follicular cells reduced cell proliferation (Stanley et al. 2012), while estrogen treatment induced proliferation and suppressed apoptosis (M. L. Lee et al. 2005). These studies suggested that these hormones have an opposite carcinogenic role in the context of the thyroid tissue.

1.3.2. Protein-level buffering of gene copy-number changes in tumours and normal cells

Cancer can be driven by somatic genetic alterations, including SNVs, CNVs, insertions and deletions (indels) and chromosomal rearrangements (Pleasance et al. 2010; Beroukhim et al. 2010; Campbell et al. 2020). A subset of these somatic alterations, known as driver mutations, confer a fitness advantage to the cancer clone and are causally implicated in oncogenesis. However, large changes in gene copy-number are known to be detrimental to the cell (Tang and Amon 2013). A plausible explanation for the detrimental phenotypes is by gene dosage alteration, where gains and losses of gene copies change the expression level of genes and proteins. The availability of measurements for gene copy-number, mRNA expression and protein abundance across a large number of tumour samples should allow us to better study the impact of gene dosage alterations at the protein level. By correlating gene copy-number changes with mRNA and protein expression measures, we can study the extent by which CNV changes are differentially propagated to the mRNA and protein levels.

Gene dosage alteration can have a negative impact on the overall cellular fitness because (i) protein overexpression can cause an energetic stress and overload protein quality-control systems, such as chaperone-mediated protein folding and the ubiquitin-proteasomal protein degradation pathway (Olzscha et al. 2011); (ii) imbalances on the stoichiometry of protein complexes can contribute to proteotoxicity because of the unbound protein subunits (Kaizu, Moriya, and Kitano 2010; Vavouri et al. 2009). Such negative effects on basic cellular mechanisms of protein maintenance and clearance can lead to protein aggregates, which are inherently toxic and the cause of many diseases (Bucciantini et al. 2002; Soto and Pritzkow 2018). Furthermore, this biology fits with the observation that

germline CNVs are rare and deleterious, and, therefore, under negative selection in the human population (Itsara et al. 2008).

Studies in aneuploid yeast strains and human cells have shown that mRNA and protein expression largely scale with most autosomal gene duplications, with the notable exception of protein complex members, which tend to be attenuated at the protein level. In other words, subunits of protein complexes show less protein abundance than expected (Pavelka et al. 2010; Dephoure et al. 2014; Ishikawa et al. 2017; Stingele et al. 2012). It was further shown that post-transcriptional mechanisms of gene dosage compensation can attenuate the protein excess through ubiquitin-dependent degradation (Dephoure et al. 2014; Ishikawa et al. 2017). In cancer the story is similar, as it has been observed that a large fraction of somatic CNVs is attenuated at the protein level (Geiger, Cox, and Mann 2010; Gonçalves et al. 2017). In addition, some protein complex members can act as scaffolding or rate-limiting for the assembly of the complex, indirectly controlling the degradation level of attenuated complex subunits that depend on their binding state, i.e., bound/unbound to the complex (Gonçalves et al. 2017). These results are lined up with pulse-chase degradation experiments showing that several protein complex subunits are degraded at a higher rate when unbound from the corresponding complexes (McShane et al. 2016). Post-transcriptional regulation of protein expression, likely via protein interactions and degradation, also explains the cancer-associated events of collateral loss, or reduced protein expression, in protein complexes following mutation in one subunit (Ryan et al. 2017; Roumeliotis et al. 2017), and why correlation analysis can be used to find cancer-specific protein interaction networks (Lapek et al. 2017).

Degradation of unbound protein complex subunits may happen to avoid free hydrophobic interface sites that can be prone to aggregate (Young, Jernigan, and Covell 1994) and supports a long-established view that protein complex formation can set the total amount of protein levels (Abovich et al. 1985). Despite the events of gene dosage compensation by post-transcriptional regulation have been well documented in cancer, we still do not understand (i) what protein properties are associated with the propensity for a protein to be attenuated and (ii) if the characteristics of the attenuation process are also seen in noncancerous cells.

In chapter **3**, a multi-omics study of protein-level attenuation of gene dosage was carried out to address these questions and explore further the events of post-transcriptional regulation in cancer.

1.3.3. Variability of kinase and transcription factor activities in cancer

Mutations in key cancer genes are just the first steps of a cascade of events that culminate in neoplastic disease and eventually cancer. In reality, these mutations generate the genetic diversity that promotes the acquisition of multiple cancer hallmarks, including sustaining chronic proliferation; resisting cell death; inducing angiogenesis; escaping from immune surveillance; and activating tissue invasion and metastasis (Hanahan and Weinberg 2011). One of the most fundamental traits of cancer cells is the ability to sustain proliferative signalling. A cancer cell can acquire this capability in a number of alternative ways: (i) producing growth factor ligands itself or stimulating the normal cells in the tumour stroma to produce them; (ii) increasing the amount of receptor proteins in the cytoplasmic membrane; (iii) by constitutive activation of signalling transduction pathways due to mutations on their components (Hanahan and Weinberg 2011). In fact, perturbation of kinase signalling by genetic alterations often results in deregulated kinase activity and malignant transformation (Blume-Jensen and Hunter 2001). As an example, about 40% of melanomas contain the V600E activating mutation in the BRAF kinase, resulting in constitutive signalling through the Raf to mitogen-activated protein kinase (MAPK) pathway and increased cellular proliferation (Davies and Samuels 2010).

Aberrant kinase signalling in cancer has been studied for the past two decades (Blume-Jensen and Hunter 2001; Dhillon et al. 2007; J. Yang et al. 2019). Early MS-based studies in lung cancer cell lines and human embryonic kidney cells have identified tyrosine phosphorylation signalling pathways that are activated by mutated and/or overexpressed oncogenes such as EGFR, KRAS and SRC (Guo et al. 2008; Guha et al. 2008; Rikova et al. 2007; Amanchy et al. 2008). More recently, Creixell et al. performed a comprehensive study of how mutations can affect signalling pathways in ovary cancer cell lines. They established three processes whereby mutations can perturb signalling networks, including the dysregulation of signalling network dynamics by constitutive activation or inactivation of kinases; changes in the network structure by rewiring upstream or downstream interactions; and the modulation of molecular logic gates by creation or destruction of phosphorylation sites (Creixell et al. 2015). Another study found that cancer mutations close to tyrosine phosphorylation sites can induce molecular switches that alter signalling networks (Lundby et al. 2019).

Multiple studies from CPTAC (for details about this consortium please refer to subchapter **1.2**) have identified differentially phosphorylated phosphosites associated with mutations in oncogenes and TSGs. In breast cancer, 62 phosphosites were up-regulated in *PIK3CA*-mutated tumours, with 58% of them exhibiting increased PI3K pathway activity

(Mertins et al. 2016). In HBV-related liver tumours, the phosphorylation of ALDOA on S36 was significantly higher in *CTNNB1*-mutated tumours, which may drive the glycolytic metabolism and strongest proliferation of these tumours (Gao et al. 2019). Twelve lung tumours contained *KEAP1* mutations that increased the phosphorylation of NFE2L2 on S215 and S433 (Gillette et al. 2020). NFE2L2 oncogenic signatures are associated with antioxidant responses, which protect cancer cells (Taguchi and Yamamoto 2017). In endometrial cancer, truncating mutations in *TP53* were associated with increased phosphorylation of PLK1 on T210 and high protein levels of 14 mitotic markers. Accordingly, increased phosphorylation of PLK1 T210 has been shown to trigger recovery from the G2 DNA damage checkpoint (Macůrek et al. 2008; Paschal, Maciejowski, and Jallepalli 2012) and mitotic entry (Vigneron et al. 2018). The brain and colon cancer studies have characterized the activation profile of selected kinases, including the cell cycle-related kinases CDK1 and CDK2 (Petralia et al. 2020; Vasaikar et al. 2019).

Transcription factors (TFs) are proteins that play a central role in the control of gene expression by influencing RNA polymerase activity in a gene-specific manner. To regulate the expression of their target genes, TFs can either directly bind to the DNA or interact with specific cofactor proteins (Bhagwat and Vakoc 2015) . Dysregulation of TF activities is common across many forms of human cancer (T. I. Lee and Young 2013). In tumours, genes encoding TFs are often affected by gain- or loss-of-function mutations that drive tumorigenesis and confer a selective advantage to the cancer cells. As prominent examples, the tumour suppressor gene TP53 and the proto-oncogene MYC are among the most commonly altered genes in cancer (E. Y. H. P. Lee and Muller 2010; Bretones, Delgado, and León 2015). Furthermore, oncogenic kinase signalling can alter the activity of downstream TFs and implement gene expression changes that drive malignant transformation (Darnell 2002). The diversity of oncogenic mechanisms driving TF dysregulation highlights the importance of aberrant gene expression in cancer and justifies the consideration of TFs as potential targets of anti-cancer drugs (Bhagwat and Vakoc 2015; Bushweller 2019). Previous studies have used different strategies to quantify the impact of oncogenic mutations on the activities of TFs and evaluate their potential as markers of drug response (Alvarez et al. 2016; Osmanbeyoglu et al. 2017; Garcia-Alonso et al. 2018). Despite all of these efforts, a systematic Pan-Cancer analysis of the impact of mutations on kinase and TF activities is lacking. Such study is described throughout chapter **4**.

1.4. Genomics

The following subchapters cover the omics fields that are more important to the cancer questions addressed in this thesis. This subchapter is focused on cancer genomics, whereas the subchapters **1.5**, **1.6** and **1.7** discuss the main technological and methodological details of transcriptomics, proteomics and phosphoproteomics, respectively.

A genome is the complete set of DNA molecules in an organism. It provides all of the information required by the organism to function under homeostatic conditions or to respond to environmental insults. In humans and all eukaryotes, the DNA is tightly coiled around proteins called histones and organized into chromosomes, which are stored in the cell nucleus. At a lower level, the DNA is further subdivided into smaller sections called genes, which code for RNA and protein molecules. Genomics is the study of the function, structure and evolution of genomes (Sherman and Salzberg 2020).

The Human Genome Project (HGP) started in 1990 and was a 13-year long effort to obtain the first human genome sequence, costing more than $2 billion US dollars over this period (Sherman and Salzberg 2020). The HGP was accomplished with Sanger sequencing, a chain-termination method described in 1977 by Edward Sanger and colleagues (Sanger, Nicklen, and Coulson 1977). The conclusion of the HGP encouraged the development of cheaper and faster sequencing methods, resulting in the establishment of the second-generation sequencing or NGS technologies (Shendure et al. 2017). NGS platforms perform massive parallel DNA sequencing, during which millions of fragments of DNA from a single sample are sequenced in parallel, allowing an entire human genome to be sequenced in about a day. Nowadays, a common NGS approach is the short-read sequencing by Illumina. In this approach, DNA molecules are sequenced by synthesizing a complementary strand of DNA using fluorophore-labelled nucleotides (Goodwin, McPherson, and McCombie 2016). NGS technologies became an indispensable tool for cancer genomics, as described in the following subchapters.

1.4.1. Overview of cancer genomics

Cancer is an heterogeneous disease caused by the acquisition of genetic alterations in a healthy cell's genome. A myriad of genetic changes in key genes allow a healthy cell to aberrantly co-optimize multiple cellular pathways and ignore normal constraints and cell

cycle checkpoints. Such optimizations usually induce malignant transformation, autonomous expansion of the mutant cell and spread of the cancer clone (Campbell et al. 2020). NGS technologies have enabled the systematic characterization of the genetic variation at the whole-genome or exome scale (Pleasance et al. 2010), and supported the establishment of previously mentioned cancer genomics consortia (i.e., ICGC and TCGA). From the clinical point of view, NGS is being increasingly used in oncology to guide the diagnosis, prognosis and care of cancer patients (Berger and Mardis 2018). A typical cancer genome analysis pipeline includes the following steps: DNA sequencing of tumour-normal pairs, alignment of the sequencing reads to a reference genome, variant calling and, lastly, variant annotation (Vazquez, de la Torre, and Valencia 2012; Koboldt 2020) (**Figure 1.3**).

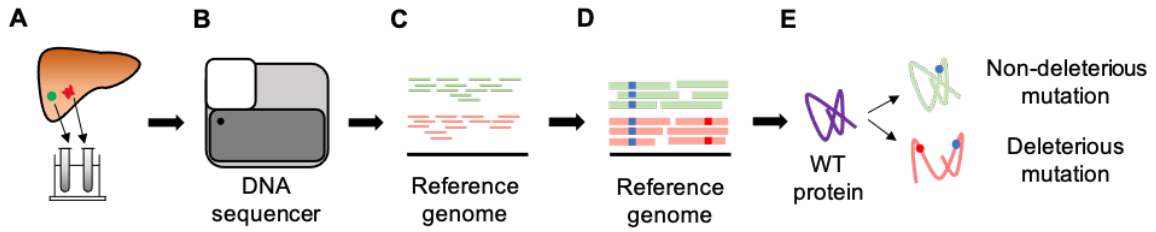


Figure 1.3. Typical cancer genomics workflow. (A) Collection of tumour and non-malignant samples (controls) from the same patient. Non-malignant samples are usually collected from the tissue-of-origin and/or patient's blood. **(B)** DNA sequencing of the tumour and normal samples. **(C)** Alignment of the tumour- (red) and normal-derived (green) sequencing reads to the reference human genome. **(D)** Variant identification is accomplished by comparing the sequencing reads to the reference genome sequence. The non-malignant control is important in this step to identify tumour-specific somatic mutations. **(E)** The variant calls are interpreted using the genome annotation to identify mutations that may alter the function of the resulting proteins. The impact of mutations on protein function can range from irrelevant to highly deleterious. WT protein: wild type protein.

1.4.1.1. Tumour sequencing

NGS of tumour specimens starts usually with the sequencing of DNA from a tumour sample and a matched control sample from the same patient (**Figure 1.3A, Figure 1.3B**). Although about 10% of cancer patients harbour germline variants that predispose to cancer (K. lin Huang et al. 2018), the main purpose of tumour sequencing is often the identification of somatic mutations with clinical relevance. The inclusion of patient-derived non-malignant DNA is crucial to distinguish somatic mutations from inherited germline variants. Non-malignant DNA samples can be obtained from the patient's blood, skin cells and fibroblasts. When non-malignant DNA is not available, mutations identified in the tumours must be

prioritized according to databases of recurrent somatic and germline variants (Hiltemann et al. 2015).

1.4.1.2. Read alignments

After the samples have been sequenced, the sequencing reads are aligned to a reference genome sequence (**Figure 1.3C**). Alignment algorithms seek to identify the original location of the sequenced fragments resulting from the tumour and control libraries. A common procedure is the use of the Burrows-Wheeler Alignment tool (BWA) paired with the Sequence Alignment/Map (SAM) format (BAM for the binary version) to store the read alignments against the reference sequences (H. Li and Durbin 2009; H. Li et al. 2009). The quality of the alignments can be improved with multiple data pre-processing steps: data cleaning operations such as the removal of PCR artifacts and duplicated reads are crucial to increase the accuracy of the downstream variant calling algorithms. These tasks can be performed using Picard and Sambamba (Tarasov et al. 2015). Additionally, according to the recommended best practices (e.g., GATK (McKenna et al. 2010; Van der Auwera et al. 2013)), base quality score recalibration (BQSR) is often required.

1.4.1.3. Variant calling

After reads are aligned to the reference genome, variant calling algorithms can be applied to identify different types of mutations (**Figure 1.3D**). These include: SNVs, small indels, CNVs, large-scale chromosomal rearrangements (amplifications, deletions, inversions, translocations, etc.) and aneuploidies. Several somatic mutation callers have been developed for this purpose, such as MuTect2 (Cibulskis et al. 2013), MuSE (Y. Fan et al. 2016), VarScan2 (Koboldt, Zhang, et al. 2012) and SomaticSniper (Larson et al. 2012). These tools use a subtraction approach of the tumour and non-malignant data. However, there are several factors that can make the detection of tumorigenic somatic mutations very challenging. Tumour purity (proportion of cancer cells in a tumour sample) should be taken into account and estimated computationally (Smits et al. 2014). For instance, the ESTIMATE algorithm employs two gene expression signatures of infiltrating stromal and immune cells to infer their proportion in the tumour sample (Yoshihara et al. 2013). In combination with a single-sample gene set-enrichment analysis, these signatures allow to

calculate scores that reflect both the presence of each cell type in the sample and the overall tumour purity.

Variant calls are usually reported in a VCF file that organizes them according to their chromosome and position. They are often accompanied by scores measuring the sequencing quality across genomic regions.

1.4.1.4. Variant annotation

The list of somatic mutations is biologically meaningless until it is interpreted at the light of the human genome annotation. This step consists in examining the variant calls to identify mutations that may alter the function of the encoded proteins: mutations in the DNA are translated into mutations in the corresponding transcripts and proteins (Vazquez, de la Torre, and Valencia 2012). SNVs can have several impacts on the amino acid sequence of the resulting protein, and consequently on the cancer cell phenotype. They can: (i) leave the amino acid sequence unperturbed (synonymous mutations); (ii) cause an amino acid substitution (nonsynonymous mutations); (iii) introduce a premature stop codon in the amino acid sequence (nonsense or truncating mutations); (iv) remove stop codons and cause read-through mutations. Indels can preserve the reading frame if they comprise a multiple of three nucleotides or cause a frameshift mutation otherwise. Large DNA rearrangements that connect coding exons from different genes are interpreted using the existing annotation for those exons to generate the amino acid sequence of the fusion protein (Mardis 2018).

The impact of the amino acid alterations on protein function can range from neutral to highly deleterious (**Figure 1.3E**). Neutral mutations do not significantly affect the structure and stability of the proteins, or do not affect regions involved in biochemical reactions and protein-protein interactions. On the other hand, deleterious amino acid changes can result in the formation of a truncated protein lacking important functional regions (Vazquez, de la Torre, and Valencia 2012). These alterations can be predicted using specialized software tools such as the Ensembl VEP (McLaren et al. 2016), ANNOVAR (K. Wang, Li, and Hakonarson 2010) and Oncotator (Ramos et al. 2015).

The mutation types discussed so far are called coding mutations because they affect protein-coding genomic sequences (i.e., exonic regions, see subchapter **1.5** for details). However, more than 98% of the human genome does not encode proteins (Garraway and Lander 2013), suggesting that the vast majority of cancer mutations are within noncoding regions (introns, promoters, enhancers, etc.). According to cancer statistics from the

Catalogue Of Somatic Mutations In Cancer (COSMIC), the actual knowledge about cancer mutations is largely biased for coding regions (Tate et al. 2019). This is mostly due to the fact that cancer genome sequencing studies have largely focused on the exome rather than on the whole genome for reasons of cost (Khurana et al. 2016). Noncoding mutations are by definition more challenging to study and interpret. As they do not affect coding regions directly, they may alter the gene expression of target genes and respective protein abundances by affecting regulatory mechanisms. A strategy that has been used to interpret their functional roles is to overlap the noncoding mutations with DNA regulation data from the Encyclopedia of DNA Elements (ENCODE) (Dunham et al. 2012). Another alternative is to perform eQTL studies to link mutated loci in noncoding regions to the expression of putative target genes (W. Zhang et al. 2018).

1.4.2. Mutagenic processes and signatures in cancer

DNA damage occurs constantly: it is estimated that the average human cell receives tens of thousands of DNA lesions per day, most of which are efficiently repaired by DNA repair pathways (Saul and Ames 1986; Sancar et al. 2004; Lindahl and Barnes 2000). Apart from being the fuel of evolution and natural selection, DNA damage is involved in many processes less beneficial to an organism. When DNA lesions are not repaired or are repaired incorrectly, they lead to mutations or wider-scale genome aberrations that threaten the cells and the organism viability (Jackson and Bartek 2009). In fact, DNA damage is implicated in ageing (Hoeijmakers 2009) and cancer (Stratton, Campbell, and Futreal 2009).

There are exogenous and endogenous sources of DNA damage that can promote cancer (Tubbs and Nussenzweig 2017). The most common exogenous genotoxins implied in human cancer are ultraviolet (UV) light (Armstrong and Kricker 2001), ionising radiation (Gilbert 2009), biological agents (viruses and bacteria) (Zur Hausen 1991) and various chemicals, such as alkylating agents (D. Fu, Calvo, and Samson 2012) and polycyclic aromatic hydrocarbons (Mastrangelo, Fadda, and Marzia 1996). Endogenous sources of DNA damage that may contribute to mutations in cancer genomes include: inactivation of DNA repair pathways by loss of DNA repair genes (Nowell 1976; Kinzler and Vogelstein 1997); stochastic errors in DNA replication caused by misincorporation of nucleotides by DNA polymerases (Tomasetti and Vogelstein 2015; Alexandrov and Stratton 2014); and oncogene-induced DNA replication stress followed by error-prone repair and genomic instability (Halazonetis, Gorgoulis, and Bartek 2008; Negrini, Gorgoulis, and Halazonetis

2010). According to this hypothesis, activated oncogenes induce the stalling and collapse of DNA replication forks, which in turn leads to formation of DNA double-strand breaks.

High-throughput DNA sequencing flooded cancer genetics with an enormous amount of data that quickly revealed a remarkable diversity of somatic mutations, as well as cancer-specific mutational signatures of DNA damage and errors in DNA repair. For instance, in melanoma it was found a high proportion of cytosine to thymine mutations (C>T) at adjacent pyrimidine nucleotides, a mutational spectrum that has been previously associated with UV light exposure (Pleasance et al. 2010). Tobacco smoking was also associated with a higher rate of specific signatures in smokers versus nonsmokers in different cancer types (Alexandrov et al. 2016). More recently, the PCAWG consortium provided a systematic characterization of mutational signatures associated with exogenous and endogenous exposures, across more than 20,000 cancer genomes from the TCGA and ICGC consortia (Alexandrov et al. 2020).

1.4.3. Cancer genes and driver mutations

Most tumors harbor a constellation of somatic genomic alterations that include SNVs, indels, CNVs and large DNA rearrangements (Garraway and Lander 2013). These mutations are especially harmful when they hit two distinct classes of cancer-causing genes: proto-oncogenes and TSGs (Garraway and Lander 2013; E. Y. H. P. Lee and Muller 2010) . Proto-oncogenes typically encode proteins that stimulate cell division and inhibit cell death. Thus, mutations in proto-oncogenes are usually classified as gain-of-function (or activating mutations) and promote cell growth, division, and survival. The mutated version of a proto-oncogene is called an oncogene. Classical oncogenes are *EGFR*, *KDR*, *HRAS* and *KRAS*. On the other hand, TSGs often restrain inappropriate cell growth and division, and are involved in DNA repair processes that prevent the accumulation of mutations in cancer-related genes. Therefore, mutations in TSGs are normally classified as loss-of-function (or inactivating mutations) and inactivate DNA repair and cell cycle control processes. Examples of TSGs include *TP53*, *RB1*, *APC*, *BRCA1/2* and *PTEN*. Another important aspect in cancer genomics is the distinction between driver and passenger mutations. Driver mutations confer a fitness advantage to the cancer clone and are therefore directly implicated in tumorigenesis (Martínez-Jiménez et al. 2020). These mutations are under positive selection and are recurrent (i.e., highly frequent) across the cancer cohorts. The majority of them tend to affect TSGs such as *TP53* (Rivlin et al. 2011) or oncogenes like *KRAS* (M. T. Wang et al. 2015). Passenger mutations do not affect the fitness of the

cancer cells but are carried by the same cells which have acquired the driver mutations. Passenger mutations tend to accumulate over the course of tumour growth, complicating the identification of true driver mutations (Gonzalez-Perez, Deu-Pons, and Lopez-Bigas 2012).

1.5. Transcriptomics

Transcriptomics is the study of the transcriptomes, which correspond to the collection of all coding and non-coding RNA molecules, or transcripts, expressed from the genome of a given cell, tissue or organism (Z. Wang, Gerstein, and Snyder 2009). In contrast with the genome, which is usually stable, the transcriptome is by nature cell dependent (Kolodziejczyk et al. 2015; Bach et al. 2017; Messmer et al. 2019; Melé et al. 2015). This is the result of a wide range of mechanisms of pre and post-transcriptional control of gene expression. Pre-transcriptional regulation includes DNA methylation, chromatin modifications and differential regulation of transcription initiation. Post-transcriptional control of gene expression occurs through modulation of RNA processing, degradation and translation rates (Greenberg and Bourc'his 2019; Klemm, Shipony, and Greenleaf 2019; Andersson and Sandelin 2020; Baralle and Giudice 2017; Schoenberg and Maquat 2012). While the coding RNAs carry information for protein synthesis, the non-coding RNAs are involved in the control of gene expression via epigenetic modifications and post-transcriptional regulation (Holoch and Moazed 2015; Jonas and Izaurralde 2015). For clarity, this thesis is focused on the coding transcriptome, since it is the most relevant for the studies here described.

To generate the coding transcriptome, the protein-coding genes are first transcribed into precursor messenger RNAs or pre-mRNAs (**Figure 1.4**). The pre-mRNAs contain protein-coding nucleotide sequences interrupted by non-coding segments, called exons and introns, respectively. In order to become functional, the pre-mRNAs undergo a series of modifications, known as mRNA processing, in the cell nucleus. These modifications increase the stability of the mRNAs and promote their translation. An important modification that is worth mentioning is RNA splicing. During splicing, the introns are removed and the coding exons are spliced together in the final mature mRNA. Remarkably, a single protein-coding gene can produce multiple transcripts, or isoforms, by alternative RNA splicing. In this process, the exons can be joined through multiple combinations, allowing a single gene to express different isoforms that may encode related proteins with different functions. After

processing, the mRNA is transported to the cytoplasm and translated into proteins by the ribosomes (Hartwell et al. 2011).

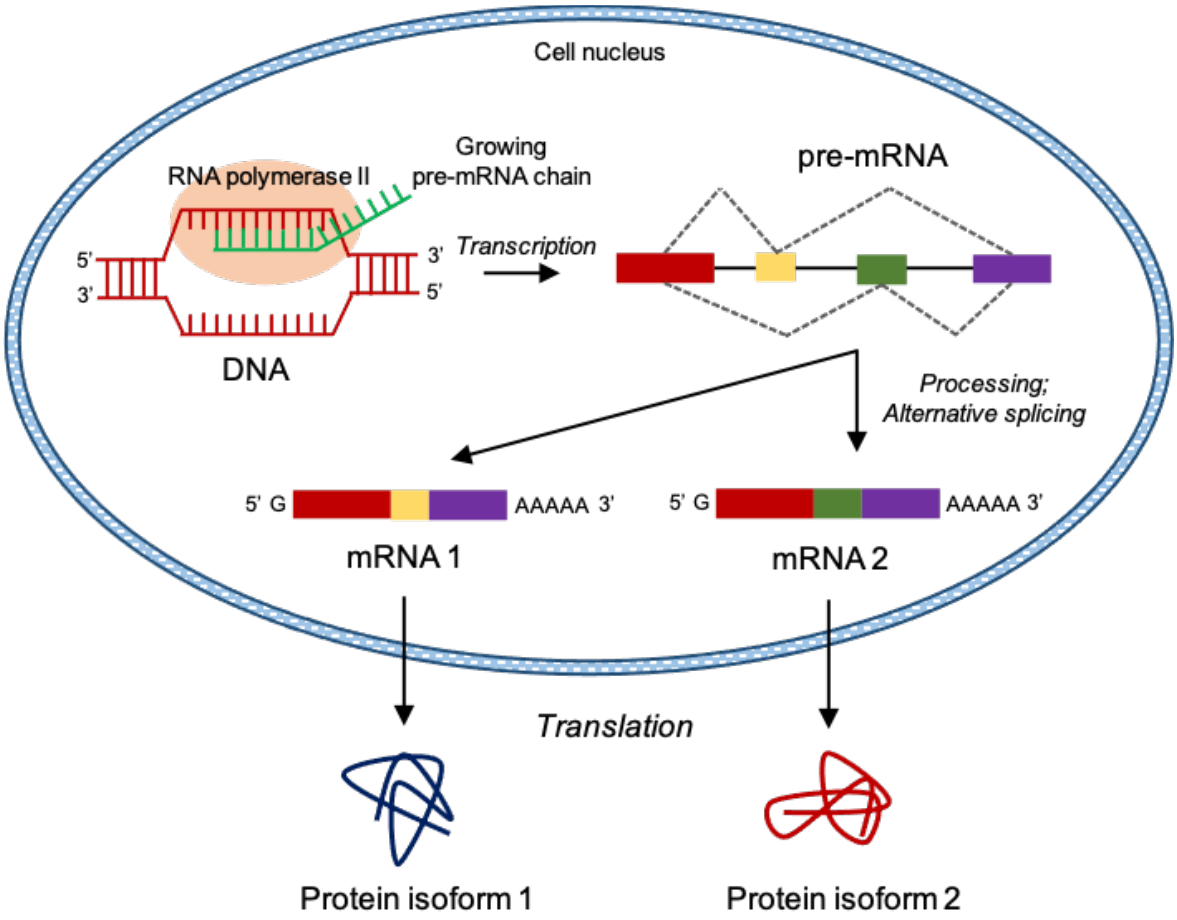


Figure 1.4. Overview of eukaryotic transcription. In the cell nucleus, a pre-mRNA is synthesized from DNA by a process called transcription. RNA polymerase II catalyses this reaction, building a complementary chain of nucleotides in the 5' to 3' direction. Eukaryotic pre-mRNAs are composed of coding exons (coloured rectangles) and non-coding introns (solid black lines). After transcription, the pre-mRNAs are modified in a process called processing. The 5' and 3' ends of the pre-mRNA molecule are connected to a 7-methylguanylate group (cap) and a poly-A (adenine nucleotides) tail, respectively. These modifications increase the stability of the molecule, prevent its degradation and promote mRNA translation. pre-mRNA processing also encompasses RNA splicing: the introns are removed and the exons are connected in the final mRNA. During alternative splicing, the exons can be joined in different combinations, yielding different mRNA molecules called isoforms. The information carried by the mRNA is then translated into proteins, which make up the structure of cells and are responsible for most of its functions. Translation occurs in ribosomes in the cytoplasm with the intervention of two other types of RNA molecules: transfer RNA (tRNA) and ribosomal RNA (rRNA). tRNA transports the amino acids to the growing polypeptide chain and rRNA is a component of ribosomes. Alternative splicing enables a single gene to express different protein isoforms.

Several large scale transcriptomic studies aim to quantify the variability of gene expression in healthy and disease tissues across a large number of individuals (Melé et al. 2015; Ferreira et al. 2014). The progresses made in the NGS technologies established the RNA-seq (RNA sequencing) as the standard method for transcriptome-wide analysis (Z. Wang, Gerstein, and Snyder 2009). This method is addressed in the next subchapter.

1.5.1. RNA-seq: sequencing the transcriptome

In RNA-seq, transcripts are identified and quantified using deep-sequencing technologies. This work is focused on the RNA-seq method most commonly used by the scientific community, termed bulk RNA-seq, which consists in the characterization of the transcriptome of groups of cells or tissue sections.

Short-read sequencing by Illumina has been the dominant technology for bulk RNA-seq. The typical Illumina short-read sequencing workflow includes RNA extraction and fragmentation (**Figure 1.5A-Figure 1.5C**), cDNA synthesis (**Figure 1.5D**), adaptor ligation (**Figure 1.5E**), PCR amplification (**Figure 1.5F**), cDNA sequencing (**Figure 1.5G**) and data analysis (**Figure 1.5H-Figure 1.5J**). During sequencing, the cDNA library is clustered in a flow cell and the individual cDNA molecules are sequenced by synthesizing a complementary strand of DNA with fluorescently labelled nucleotides. In each sequencing round, the growing strand of DNA is scanned by the sequencer to detect which of the four nucleotides has been added. This technology generates reads with 50-500 base pairs (bp) that represent fragments of the cDNA molecules. Importantly, in short-read RNA sequencing the reads can be obtained from one end or both ends of the cDNA molecules, establishing the so-called single-end or paired-end reads, respectively. The paired-end reads are separated by a known distance, allowing the alignment algorithms to connect exons across long ranges and to resolve more efficiently the multiple splicing isoforms of a single gene. The cDNA libraries are usually sequenced to an average read depth of 20-30 million reads per sample (R. Stark, Grzelak, and Hadfield 2019).

The data analyses usually involve 4 different steps, including (i) alignment and assembly of the sequencing reads (**Figure 1.5H**); (ii) quantification of gene and/or transcript abundance (**Figure 1.5H**); (iii) normalization and filtering of the expression matrix (**Figure 1.5I**); (iv) gene functional analyses and data mining (**Figure 1.5J**). The following subchapters address each of these steps.

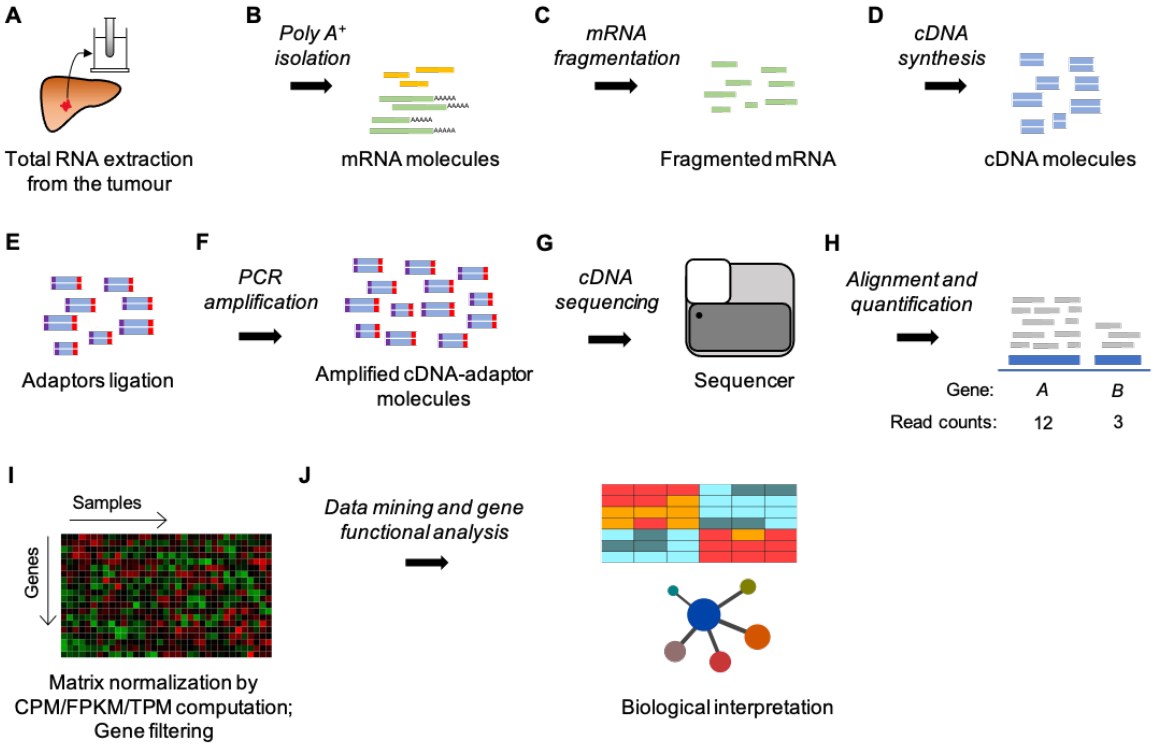


Figure 1.5. Cancer transcriptomics analysis pipeline. (A) A standard RNA-seq experiment starts with RNA extraction from the tumour tissue. Depending on the biological question, it may be advisable to extract RNA from non-malignant tissue to have a normal control. **(B)** Messenger RNAs (mRNAs) are usually selected by isolating the polyadenylated (poly A+) RNAs. **(C)** mRNAs are then fragmented using RNA hydrolysis or nebulization. **(D)** The RNA fragments are reverse transcribed into cDNA molecules. **(E)** Sequencing adaptors (purple and red) are connected to the cDNA molecules. **(F)** The cDNA library is amplified by polymerase chain reaction (PCR). **(G)** The amplified cDNA library is finally sequenced using NGS technologies to produce millions of short reads. **(H)** After sequencing, reads are pre-processed by removing low-quality reads and artefacts, such as adaptor sequences, contaminant DNA and PCR duplicates. Next, the pre-processed reads are aligned to the reference genome. The expression level of each gene/transcript is then estimated by counting the number of reads that align to each gene/transcript. **(I)** The read counts are normalized by calculating expression units such as counts per million (CPMs), fragments per kilobase of exon per million reads mapped (FPKMs) and transcripts per million (TPMs). After normalization, lowly expressed genes are often removed to improve the results of downstream analysis. **(J)** Bioinformatic analyses are carried out in the normalized matrix to translate the gene expression data into biological findings. Figure adapted from (Martin and Wang 2011).

1.5.1.1. Alignment and assembly of the sequencing reads

Once the sequencing has been completed, the reads, usually contained in FASTQ files, are aligned to a reference annotated genome and converted into genomic coordinates (**Figure 1.5H**). The alignment can be done using different tools, such as TopHat (D. Kim et

al. 2013), STAR (Dobin et al. 2013) and HISAT (D. Kim, Langmead, and Salzberg 2015). As the reads might encompass the splice junctions between adjacent exons in the cDNA molecules synthesized from the mRNA (where an intron has been removed), these tools perform spliced alignments by allowing gaps in the reads when aligned to the reference genome (with introns and exons). When a high-quality reference genome is not available (as for non-model species) or the focus is on the transcripts rather than on the genes (e.g., finding novel aberrant transcripts in tumours), a *de novo* transcriptome can be assembled directly from the reads using, for example, the StringTie (Pertea et al. 2015) and SOAPdenovo-Trans (Y. Xie et al. 2014) assemblers.

1.5.1.2. Quantification of gene abundance

After the reads have been mapped to the reference genome, the next steps consist in assigning the reads to genes or transcripts and determining abundance measures. There are different tools that were designed to accomplish these tasks, including HTSeq (Anders, Pyl, and Huber 2015), featureCounts (Liao, Smyth, and Shi 2014), RSEM (B. Li and Dewey 2011), CuffLinks (Trapnell et al. 2012) and MMSeq (Turro et al. 2011). The quantification of genes is performed by counting the reads that are aligned to their genomic position (**Figure 1.5H**). On the other hand, the quantification of the multiple spliced isoforms is more complex because the reads that do not align to splice junctions can not be unambiguously assigned to specific transcripts (multi-mapped reads). This problem is usually solved by estimating the contribution of the multi-mapped reads across the different isoforms, using, for example, the expectation-maximization algorithm implemented in RSEM.

1.5.1.3. Normalization and filtering procedures

The number of reads mapped to a given gene in a given sample is proportional not only to the expression level of the gene, but also to its length (sum of all exon lengths) and to the total number of reads sequenced in that sample (library size). In other words, one would expect more reads from longer genes and from samples that have been sequenced to a higher depth. Therefore, the next step in RNA-seq data analysis is usually the normalization of the gene expression matrix in order to account for such differences (**Figure 1.5I**). The expression units more commonly used to report the gene expression data are the

counts per million (CPM), fragments per kilobase of exon per million reads mapped (FPKM) and transcripts per million (TPM) (Law et al. 2014; Mortazavi et al. 2008; Trapnell et al. 2010; B. Li and Dewey 2011). All these units are calculated from the number of reads mapped to a particular gene i (X_i) and from the library size (N). FPKM and TPM also take into account the gene length in base pairs (L_i). The CPM of gene i is calculated from the following formula:

Equation 1.1: $CPMi = \frac{Xi}{\frac{N}{10^6}} = \frac{Xi}{N} \times 10^6$

CPM corresponds to the gene read counts (X_i) divided by the library size (N) and multiplied by one million. This is a simple unit that is used by the differential gene expression method limma (Law et al. 2014). The FPKM of gene i can be calculated as follows:

Equation 1.2: $FPKMi = \frac{CPMi}{\frac{Li}{10^3}} = \frac{Xi}{\frac{Li}{10^3} \times \frac{N}{10^6}} = \frac{Xi}{Li \times N} \times 10^9$

FPKM corresponds to the CPM divided by the gene length in kilobases ($L_i/10^3$). The interpretation is the following: if that RNA sample was sequenced again, it would be expected to see, for gene i, $FPKM_i$ fragments per million reads sequenced and per thousand bases in the gene. The TPM for gene i can be calculated using the following equation:

Equation 1.3: $TPMi = \frac{Xi}{Li} \times \frac{1}{\Sigma j \frac{Xj}{Lj}} \times 10^6 = \frac{FPKMi}{\Sigma j FPKMj} \times 10^6$

TPM corresponds to the read counts per base (X_i/L_i) divided by their sum across all genes ($\sum_i X_j/L_j$) and multiplied by one million. Alternatively, TPM can be computed as the FPKM over the sum of all FPKMs ($\sum_j FPKM_j$) multiplied by one million. TPM corresponds to the proportion, or abundance, of gene i in relation to the other genes j in the same sample. As the sum of all TPMs in each sample equals one million, this measure allows to compare the proportion of mapped reads to a given gene across samples. In contrast, the sum of all FPKMs may be different across samples, making it more difficult to compare samples directly. For this and other reasons, the scientific community has been advocating the use of TPMs over FPKMs in RNA-seq related studies (G. P. Wagner, Kin, and Lynch 2012).

None of these units (CPM, FPKM and TPM) take into account different RNA pool compositions between samples. As an example, consider the following hypothetical

scenario: two samples A and B are sequenced to the same read depth (number of reads). Every gene that is expressed in B is also expressed in A at the same expression level (number of transcripts). However, sample A also expresses a set of genes equal in number and expression that are not expressed in B. Thus, the RNA production of sample A is twice the size of sample B. As both samples are sequenced to the same depth, a gene expressed in both samples has half the reads in sample A, since the reads are spread over twice as many genes. If the different RNA populations are not adjusted for, this bias can force the differential expression analysis methods to be skewed towards one experimental condition. The trimmed-mean of M values (TMM) was developed to normalize the library sizes for such biological confounding factors (Robinson and Oshlack 2010). This method estimates appropriate scaling factors to convert the original library sizes into effective library sizes. Importantly, TMM is a batch normalization method, which means it was developed to use in a group of samples. While the CPM, FPKM and TPM are calculated in each sample and are not affected by the other samples, the TMM scaling factors need to be recalculated if the sample set changes.

After normalization, the expression matrix is usually filtered to remove those genes with low expression, a process that has been shown to remove noise and improve the detection of differentially expressed genes (Bourgon, Gentleman, and Huber 2010).

1.5.1.4. Gene functional analyses and data mining

Once the expression matrix has been properly normalized and filtered, the gene expression data is finally evaluated in a biological context through bioinformatic analyses (**Figure 1.5J**). In transcriptomics, the most common approaches for studying gene function are through (i) differential gene expression modelling; (ii) gene correlation networks; (iii) gene set enrichment analysis; and (iv) transcription factors activities inference.

1.5.1.4.1. Differential gene expression

Differential gene expression (DGE) methods determine quantitative changes in the expression levels of genes between two or more experimental groups (e.g., tumour vs normal samples). These genes are often referred to as differentially expressed (DE). The

methods to perform DGE analyses can be classified as non-parametric and parametric (Costa-Silva, Domingues, and Lopes 2017; Soneson and Delorenzi 2013).

Non-parametric tests make minimal assumptions about the probability distributions of the gene expression data and do not fit any strict models to the data. The Mann–Whitney U test, also known as Wilcoxon rank-sum test, is a simple non-parametric statistical test that can be used to identify DE genes between two sample groups (J. Ma, Malladi, and Beck 2016). Non-parametric approaches such as SAMseq (J. Li and Tibshirani 2013) and NOISeq (Tarazona et al. 2011) are statistically more complex. NOISeq computes, for each gene, a statistic that expresses the difference between two distributions: one calculated from the absolute expression differences between the two contrasted conditions, and another one obtained by comparing pairs of samples belonging to the same condition (noise distribution). SAMseq calculates a statistic based on the Wilcoxon rank-sum test, which is averaged over several resamplings of the data. It then uses a permutation strategy to estimate false discovery rates (Soneson and Delorenzi 2013).

Parametric tests for DGE analyses assume that the gene expression data follow specific distributions. These tests also rely on fixed parameters about the distributions in order to calculate the respective statistics. The Student's *t*-test for independent groups, also known as two-sample *t*-test, is a parametric alternative to the Wilcoxon rank-sum test. The two sample *t*-test can be applied to data that follows a normal distribution. As the gene expression data is not normally distributed *per se*, it can be transformed into a normal distribution using, for example, a base 2 logarithm (log2). Methods based on linear regression models have more statistical power (probability of correctly rejecting a false null hypothesis) to detect true differential expression (Law et al. 2014). Linear regression is an approach to study associations between variables. It builds models for predicting a given response (or dependent) variable *Y* based on a single or multiple explanatory (or independent) variables *X*. In DGE analyses, the main idea is to model the expression of each gene as a linear combination of different explanatory variables. The explanatory variables, which can be qualitative or quantitative, are tested for association with the expression of the genes. Consequently, these methods also allow the researchers to consider the effects of possible confounding variables, or confounders, on gene expression, like the covariates age and gender or even technical artifacts. As an example, the expression of a given gene *i* can be modelled by a set of *n* explanatory variables using the following expression:

Equation 1.4: $E_i = \beta_0 + \beta_1 V_1 + \beta_2 V_2 + \ldots + \beta_n V_n + \varepsilon$

Where E_i is the expression of gene i, $V_1, V_2, \ldots, V_n$ represent the multiple explanatory variables and $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the respective parameters, or coefficients, to be estimated from the data. The ε is an error term and represents unmeasured variables or unmeasurable variation that is useful to predict E_i. The β_0 coefficient is called the intercept and represents the value of E_i when the explanatory variables are equal to 0. The coefficient β_j represents the average effect on E_i of a one unit increase in V_j, while holding the other independent variables fixed. Methods that implement linear regression models in order to identify DE genes include DESeq2 (Love, Huber, and Anders 2014), edgeR (Robinson, McCarthy, and Smyth 2009) and limma/voom (Law et al. 2014, 2016). However, some important mentions have to be made in relation to the linear models employed by these methods. DESeq2 and edgeR are based on generalized linear models (GLM). A GLM is a more flexible version of the standard linear regression model which, among other things, allows the distribution of the response variable to be different from the normal distribution. The GLMs used in edgeR and DESeq2 assume that the read counts are distributed according to the negative binomial distribution. Limma, on the other hand, is based on the standard linear models that were originally developed to analyze gene expression data from microarrays. To extend the standard linear models to RNA-Seq data, limma first transforms the read counts to approximate a normal distribution using the voom method. The voom transformation consists essentially in computing the log2 CPM of the read counts and estimating the mean-variance relationship (Law et al. 2014).

Once all genes have been tested for differential expression, the P-values are then corrected, or adjusted, for multiple testing. This correction is crucial because as the number of statistical tests increases (tens of thousands in DGE analyses), the chances of incorrectly rejecting the null hypothesis and obtaining false positives (type I errors) also increases. There are multiple techniques to decrease the chances of obtaining false positives. A popular approach is called Benjamini-Hochberg (BH) procedure, and it was developed to control the false discovery rate (FDR). The FDR is the proportion of false discoveries (or false positives) among all the significant tests. To elaborate, a BH-adjusted P-value of 5% implies that 5% of the significant tests will be false positives after the correction (Akalin et al. 2021).

1.5.1.4.2. Gene co-expression networks

Gene co-expression networks represent gene relationships by linking genes with similar co-expression patterns along a given set of samples or conditions. Co-expressed

genes might be tightly co-regulated, functionally related, or members of the same pathway or biological process (Melé et al. 2015; Saha et al. 2017).

Weighted gene correlation network analysis (WGCNA) is a method to build and study gene co-expression networks (Langfelder and Horvath 2008). The main goals of WGCNA are (i) build a gene co-expression network; (ii) divide the network into gene modules; (iii) relate modules to external information, including phenotypic traits; (iv) identify driver genes in the biologically interesting modules.

In weighted gene co-expression networks, genes (nodes) are connected based on their pairwise correlations (edges). These networks are therefore undirected and weighted. WGCNA can use multiple correlation methods, including the Pearson (default) and Spearman correlation coefficients. Independently of the chosen method, the correlation coefficient expresses quantitatively the magnitude and direction of the relationship between two given genes. The adjacency matrix of the network is calculated as $A_{ij} = |corr(i,j)|^{\beta}$, where $|corr(i,j)|$ corresponds to the absolute correlation between the genes *i* and *j*. The absolute correlations are further raised to the lowest β value that approximates the network of a scale-free topology. Scale-free networks are networks whose degree distribution follows a power law, and consequently have few nodes with many connections (hubs) and many nodes with few connections. In practice, raising the absolute correlation values to a power accentuates high correlations at the expense of low correlations. This approximation is important because biological networks tend to be scale-free (Barabási and Bonabeau 2003; Broido and Clauset 2019).

Gene modules, or clusters of densely interconnected genes, are identified by converting the adjacency matrix into a topological overlap matrix (TOM). TOM contains the pairwise relative interconnectedness of all genes in the network. After converting the TOM into a dissimilarity measure (1 - TOM), gene modules are identified by cutting off the branches of a hierarchical clustering dendrogram using the dynamic tree cut algorithm (Langfelder, Zhang, and Horvath 2008). This method eliminates the need of using constant height cutoff values and is more effective in complex dendrograms. The gene expression profiles of a module can be represented by its module eigengene (first principal component). Very similar modules are then merged into a single module if their eigengenes are highly correlated (> 0.85 by default).

The biological significance of a module can be identified by correlating or fitting a simple linear regression model between the module eigengene and a given phenotype. Modules with high biological significance can represent pathways or biological processes associated with the respective phenotype. Highly connected genes inside these modules, or hub genes, are natural candidates for further validation, since their expression profiles can represent that of the entire module. Hub genes can be identified using module

membership measures, including the eigengene-based connectivity (K_{ME}). The K_{ME} corresponds to the correlation between the gene expression profile and the module eigengene. As expected, hub genes tend to have high absolute K_{ME}.

1.5.1.4.3. Over representation and gene set enrichment analysis

A gene set is a collection of genes that are functionally related, such as being involved in the same biological process or pathway. Over representation analysis (ORA) is a widely used approach to determine whether known gene sets are over-represented, or enriched, in a gene list of interest, for instance a list of DE genes derived from a DGE method or a gene module from WGCNA (Dugourd and Saez-Rodriguez 2019). ORA can be performed using a hypergeometric test or a Fisher's exact test (G. Yu et al. 2015; Reimand et al. 2019).

ORA can detect enrichment in genes that show a large expression difference between two given conditions, e.g., DE genes. However, it is unable to report situations where the expression difference is small but evidenced in a coordinated way in a group of related genes. Methods based on gene set enrichment analysis (GSEA) address this limitation by evaluating if there are gene sets where the overall expression difference is more extreme than expected. In GSEA-like approaches, all genes are used and ranked based on their phenotypic differences (i.e., fold-changes, P-values, etc.). Then, given a priori defined gene set (e.g., pathway), GSEA calculates an enrichment score, which represents whether the members of the gene set are randomly distributed throughout the ranked gene list or primarily found at the top or bottom (Subramanian et al. 2005; Reimand et al. 2019).

As mentioned above, both ORA and GSEA rely on pre-defined gene sets to calculate enrichment scores. This prior knowledge is available in many different resources. Gene ontology (GO) knowledgebase is widely used and comprehends a collection of annotated terms describing molecular functions, cellular compartments and biological processes associated with specific genes. Other types of annotations, including hallmark genes from well-studied biological processes, chromosomal positions and cancer-related genes are available in databases such as MSigDB (Liberzon et al. 2011). Signalling networks and pathways can be obtained from KEGG (Kanehisa and Goto 2000) and Reactome (Fabregat et al. 2016) databases.

1.5.1.4.4. Footprint-based activities of transcription factors

The abundance of thousands of transcripts or genes can be used as the molecular signature of a biological sample. The same concept can be applied to transcription factors (TFs). The abundance of the regulatory targets of a TF can be considered footprints of the activity of the TF (Dugourd and Saez-Rodriguez 2019).

VIPER is a rank-based statistical method that can be used to estimate the activity of TFs (Alvarez et al. 2016). The fundamental idea is to compute an enrichment score using the molecular readouts from the TF targets and use it as a proxy for the activity of the respective TF. The TF targets can be obtained from multiple resources, such as Omnipath (Türei, Korcsmáros, and Saez-Rodriguez 2016) and DoRothEA (Garcia-Alonso et al. 2019).

1.6. Proteomics

In order to maintain cellular homeostasis or even follow specific cell fates, cells must respond to environmental stimuli, such as changes in temperature, oxygen levels, and differentiation signals. In general, cells respond to environmental cues by activating a cascade of signalling transduction pathways that can dramatically alter the expression of genes, and, eventually, the abundance of proteins. Even though the genetic information for a specific cellular process is encoded by the genes, proteins are the main effectors of the chemical reactions happening inside the cells. Proteins are a group of biomolecules, composed of one or multiple polypeptide chains, that perform a vast range of cellular processes, such as biochemical reactions, signalling, molecular transport and structural support. The proteome is the overall protein content of a biological system under a specific condition or cellular state (Aslam et al. 2017). Protein levels are not only dependent on the corresponding gene expression levels but also on translation and protein degradation rates. Thus, mRNA expression alone is not sufficient to predict the respective protein abundance in different scenarios (Yansheng Liu, Beyer, and Aebersold 2016). For these and other reasons, the proteome has been considered one of the most important molecular layers that needs to be profiled in order to properly understand biological systems (Aslam et al. 2017; Altelaar, Munoz, and Heck 2013). Proteomics aims to address this problem by characterizing proteomes, including not only the quantification of protein abundances but also the study of protein structures, interactions and modifications (Aslam et al. 2017).

Protein microarrays, such as RPPA, were established for high-throughput protein expression analysis (Sutandy et al. 2013). However, the expression profiling of whole proteomes using these techniques remains challenging due to the dependency on the availability and quality of protein antibodies (Akbani, Becker, et al. 2014). The recent advances on Mass Spectrometry (MS)-based proteomics have been accelerating the study of the proteomes of many different cell types and diseases (E. C. B. Johnson et al. 2020; L. Wu et al. 2013; Gholami et al. 2013; Jiang et al. 2020; Vasaikar et al. 2019; Nusinow et al. 2020). The general setup of a MS-based proteomics workflow is dissected throughout the next sessions.

1.6.1. Mass Spectrometry-based proteomics

The most widespread MS-based workflow is termed bottom-up or shotgun proteomics, in which the proteins are fragmented and analysed by MS (Mardamshina and Geiger 2017; Aebersold and Mann 2016). A less common alternative is called top-down or native proteomics. In this approach, proteins are analysed as intact entities by MS, enabling identification of precise protein isoforms (J. C. Tran et al. 2011; Smith and Kelleher 2013). The general workflow of bottom-up MS-based proteomics is illustrated in **Figure 1.6**. This workflow is also called liquid chromatography with tandem mass spectrometry (LC-MS/MS). Typically, the experiments start with protein extraction from the cells, tissues or body fluids (**Figure 1.6A**). After that, the proteins are digested into peptides using proteases such as trypsin (**Figure 1.6B**). The next steps aim to reduce the complexity of the peptide mixture either by sample pre-fractionation or enrichment (**Figure 1.6C**). Sample pre-fractionation can be done by ion-exchange chromatography and consists in separating the peptide (or protein) mixture by physicochemical properties, such as charge and hydrophobicity. Alternatively, subpopulations of peptides carrying specific PTMs, e.g., phosphorylations, can be enriched using affinity-based resins or antibody-based immunoprecipitation (**Figure 1.6C**). This is the case of phosphoproteomics, which will be discussed in detail in a subsequent subchapter (**1.7**). The pre-fractionated peptide mixture is then separated by high-performance liquid chromatography (HPLC) to further fractionate the peptides and reduce complexity (**Figure 1.6D**). The peptides eluting from the chromatography column are subsequently ionized (for example by electrospray) and introduced into the mass spectrometer (**Figure 1.6D**), which records the mass-to-charge (m/z) ratios of the precursor peptide ions and generates the MS^1 spectra (**Figure 1.6E**). Based on the ion intensities from the precursor peptides, single peptides are selected and fragmented (commonly by

collision-induced dissociation), which generates the MS/MS (or MS^2) fragment spectrum of each selected peptide (**Figure 1.6E**). The MS^1 and MS^2 m/z ratios from the precursor peptide and its fragments, respectively, are then matched against known protein databases (e.g., UniProt) to identify the peptides and the proteins they belong to (**Figure 1.6F**). The ion intensities from the mass spectra are used for protein and peptide quantification (**Figure 1.6E**). The process of data quantification in proteomics is discussed in more detail in the next subchapter. Peptide assignments are statistically validated to control for false positive identifications and assembled into proteins (**Figure 1.6G**). Finally, the list of proteins and their quantitative changes are the basis for bioinformatic analysis (**Figure 1.6H**) (Ahrens et al. 2010; Choudhary and Mann 2010; Altelaar, Munoz, and Heck 2013).

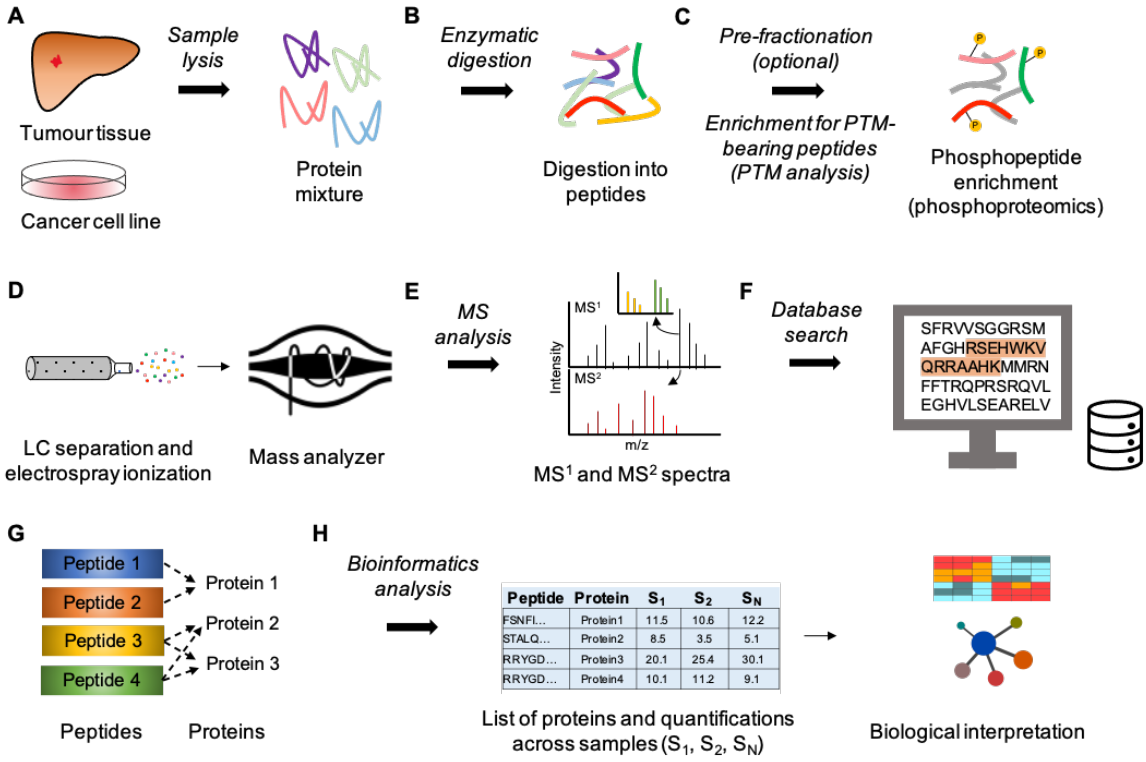


Figure 1.6. Bottom-up MS-based proteomics (LC-MS/MS) for cancer research. (A) A typical shotgun proteomics workflow starts with tumour sample lysis and protein extraction. **(B)** The protein mixture is enzymatically cleaved into peptides, using proteases such as trypsin. **(C)** Peptides are optionally pre-fractionated to reduce sample complexity or enriched for specific PTMs, for instance phosphorylations. **(D)** The pre-fractionated or PTM-enriched sample is then separated on a liquid chromatography (LC) column. Peptides eluting from the LC system are ionized by electrospray at the tip of the column and introduced into the mass spectrometer. **(E)** After ionization, peptide ions are measured in the survey-scan of the mass spectrometer, which records the MS^1 spectra of the precursor peptides. Based on their intensities, individual peptides are selected and subjected to fragmentation, which generates a characteristic fragment ion MS^2 spectrum for each precursor peptide. The ion intensities from the MS^1 and MS^2 spectra are used for relative or absolute protein quantification. The inset in the MS^1 panel represents two experimental conditions (represented by green and yellow) in which a given protein changes its abundance. **(F)** The combination of mass spectra from the precursor peptide and its

fragments are matched to large protein sequence databases to identify the peptides. **(G)** Identified peptides are assembled into proteins. This is not a trivial process because some peptides are shared between different proteins. Several models employ different strategies to identify the smallest set of proteins that explain the identified peptides. **(H)** The list of proteins and respective abundances are finally interpreted and visualized in the context of the biological question under study. Figure adapted from (Choudhary and Mann 2010; Altelaar, Munoz, and Heck 2013; Ahrens et al. 2010).

1.6.1.1. Protein quantification

The ion intensities from the MS^1 and MS^2 spectra can be used for absolute or relative quantification of peptides and proteins. Absolute quantification allows to determine the number of copies of a given protein in a sample, usually by comparing the protein signal to those of reference proteins with known quantities (spiked in standards or isotopically labeled proteins). Alternatively, absolute protein values can be estimated *in silico* from the peptide signals mapping to the protein. On the other hand, relative quantification methods provide the abundance change (or ratio) of a protein between two or more conditions (Jürgen Cox and Mann 2011).

Quantification methods can be subdivided as label-based and label-free (**Figure 1.7**). These strategies differ largely on the experimental workflows, and, consequently, on quantification accuracy and precision due to the possibility of occurrence of errors in different steps of the protocols (Bantscheff et al. 2012).

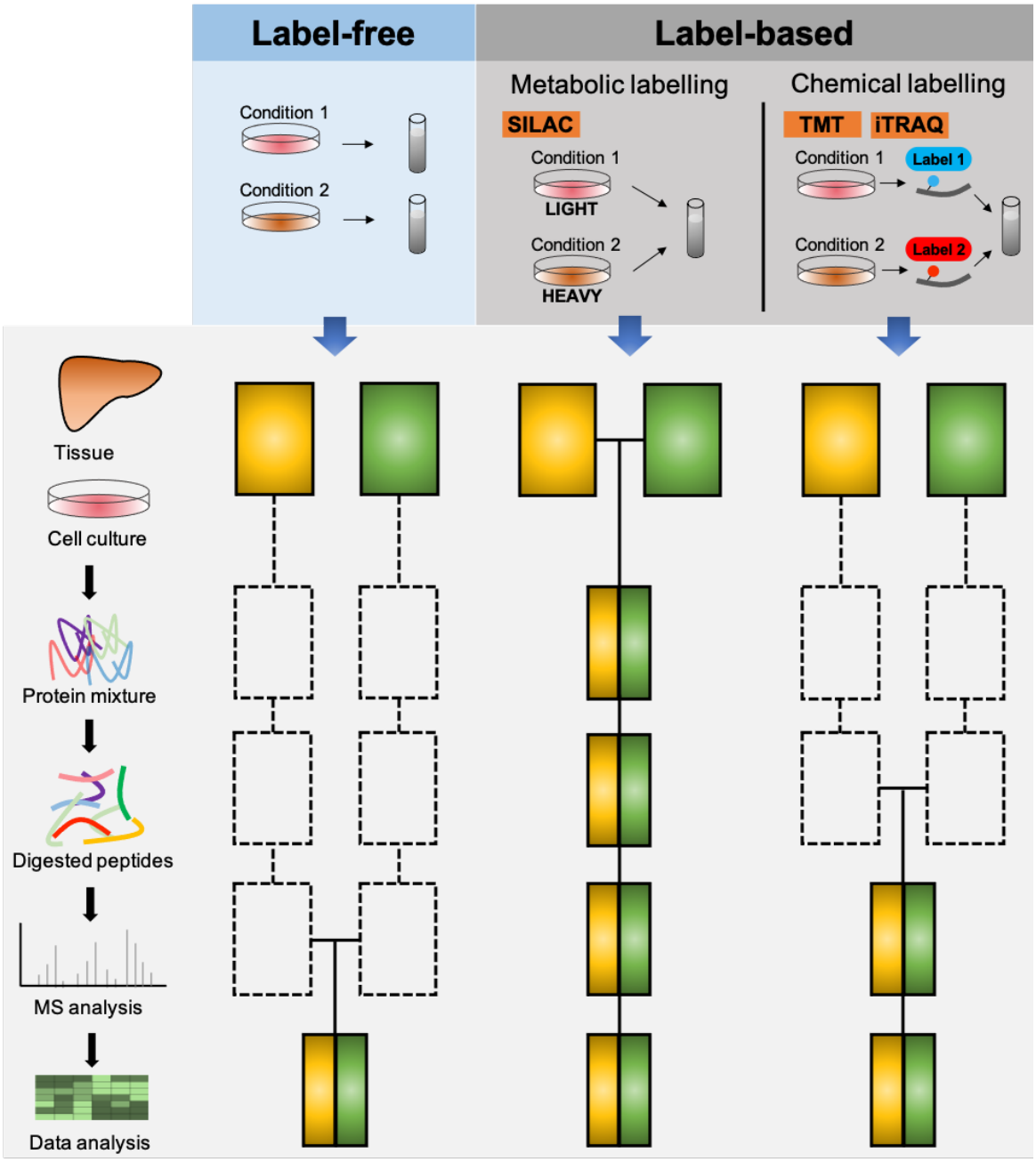


Figure 1.7. Common quantification strategies in MS-based proteomics workflows. The **top panel** represents the label-free and label-based quantitative MS protocols. Both strategies depict two different experimental conditions shown in pink and orange, although it is possible to analyse more than two samples. For instance, in cancer research the two conditions may be a tumour sample and a matched normal control. The selection of the most appropriate protocol depends on numerous factors, including the number and nature of the samples, the available mass spectrometer, and the experience to handle the samples and analyse the data. The label-free quantification does not require any extra modification to the proteomes being compared and the biological samples are analysed separately. In contrast, with label-based quantification, the proteins or peptides in each biological condition are marked differently and then mixed to be analysed in the same LC-MS/MS run. The metabolic labelling procedure involves the introduction of heavy isotopes into proteins by growing the cells with nutrients enriched with heavy atoms, e.g., the SILAC protocol uses isotopically labelled arginines and lysines. Methodologies based on isobaric tagging, like TMT and iTRAQ, label the peptides from the different samples after protein digestion. The

differences in m/z ratios between the differentially labelled proteins allow their relative quantification with the mass spectrometer. The **bottom panel** overlaps the quantification strategies with the proteomics workflow represented in **Figure 1.6** (see figure for more details). Yellow and green rectangles represent two experimental samples. The experimental point where the samples are combined is represented by horizontal lines. Dashed lines indicate experimental steps susceptible to variation and where quantification errors might occur. Figure adapted from (Bantscheff et al. 2012).

1.6.1.1.1. Label-free methods

In label-free quantification, proteins are quantified without using chemical labels and the sample lysates are not mixed (**Figure 1.7**). This means that these methods are especially useful when many experimental conditions need to be compared. Label-free strategies can be used for absolute and relative quantification of protein abundance (Bantscheff et al. 2012). Spectral counting is the simplest label-free quantification method, where the number of peptide-spectrum matches (PSMs) for a specific protein is used as a proxy of protein abundance. A more accurate approach consists in integrating the MS^1 signal intensities of the detected peptides (Jürgen Cox and Mann 2011; Bantscheff et al. 2012). Since all samples are separately analysed by LC-MS/MS, stable and comparable experimental setups, as well as more sample replicates, are required throughout the entire workflow (Bantscheff et al. 2012; Altelaar, Munoz, and Heck 2013; Choudhary and Mann 2010). Nevertheless, specialized algorithms can be used to normalize the peptide signals between different LC-MS/MS runs (J et al. 2014). In fact, several successful proteomic studies using label-free quantification have been recently reported (Lawrence et al. 2015; B. Zhang et al. 2014; E. C. B. Johnson et al. 2020).

1.6.1.1.2. Label-based methods

The differential labelling of two or more samples creates labelled peptides that can be distinguished in the MS spectra due to their distinct masses (Altelaar, Munoz, and Heck 2013; Ahrens et al. 2010). Like in label-free methods, label-based strategies can be used for absolute and relative protein and peptide quantification. The quantification of a sample of interest happens by comparing the peptide intensity against a reference labelled peptide in the same sample (absolute quantification) or against a labelled peptide from a different

sample (relative quantification) (Bantscheff et al. 2012). There are two main approaches in label-based quantification: metabolic labelling and chemical labelling (**Figure 1.7**).

In metabolic labelling, cells are cultured in a defined isotope media, incorporating the amino acid isotopes during regular protein biosynthesis (Ong and Mann 2005). Stable isotope labelling by amino acids in cell culture (SILAC) is a popular quantification technique (Ong et al. 2002). In SILAC, cells are cultured in a medium with isotopically labelled arginines and lysines (usually with ^{13}C and ^{15}N). After cell lysis and protein digestion by trypsin, it is ensured that most of the peptides contain at least one labelled amino acid (Bantscheff et al. 2012). In a typical SILAC experiment, two populations of cells are cultured in a labelled (or heavy) and normal (or light) amino acid media (**Figure 1.7**). The heavy and light amino acids have a known mass difference, allowing the proteomes to be distinguished (at the MS^1 intensity levels) and relatively quantified during MS data analysis (Choudhary and Mann 2010). A maximum of three samples are usually compared due to limitations on the number of useful heavy isotopes (Bantscheff et al. 2012). This technique has the advantage that the samples can immediately be combined after cell lysis and then be manipulated together (**Figure 1.7**). Therefore, SILAC-based quantification methods have a high quantitative accuracy due to very low variability in the experimental procedures (Jürgen Cox and Mann 2011).

The most popular chemical labelling techniques are the tandem mass tags (TMT) (Thompson et al. 2003) and the isobaric tag for relative and absolute quantification (iTRAQ) (Ross et al. 2004; Wiese et al. 2007), both of which enable to chemically label the digested peptides (**Figure 1.7**). In comparison to SILAC, these approaches are more versatile and straightforward to quantify culture-independent conditions, such as tissue samples (Altelaar, Munoz, and Heck 2013). Furthermore, quantification by isobaric tags can be multiplexed, allowing to analyse up to 16 (16-plex) different conditions in parallel in a single MS run (Thompson et al. 2019). In general, these approaches attach isobaric labelling reagents to the peptide N-terminus. Absolute and relative quantification is based on the intensities of the reporter ions in the MS^2 fragmentation spectra (Bantscheff et al. 2012; Choudhary and Mann 2010).

1.6.1.2. Computational analysis of MS data

Analysis of MS-based proteomics data encompasses different steps (Choudhary and Mann 2010; Sinitcyn, Rudolph, and Cox 2018). Initially, the raw data is pre-processed in order to quantify and identify the peptides and proteins. Once this step is concluded, one

obtains a matrix with proteins as rows and samples as columns, and protein abundances, such as ratios, in the matrix cells. This matrix is usually the starting point of bioinformatic analyses that have the ultimate goal of translating the data into biological findings. These steps are described below.

1.6.1.2.1. MS raw data pre-processing

Absolute or relative protein quantification either with the aid of labels or label-free (see subchapter **1.6.1.1**) can be accomplished using different tools including the MaxQuant computational platform (Jürgen Cox and Mann 2008; Jurgen Cox et al. 2009; Tyanova, Temu, and Cox 2016). Peptide identification from the fragmentation spectra can be performed by database search algorithms or *de novo* sequencing. *De novo* sequencing methods infer the peptides sequences directly from the fragmentation spectra, relying on the previous knowledge that known mass differences between peaks correspond to certain amino acids. This method can be performed using tools such as PEAKS (B. Ma et al. 2003), SPIDER (Han, Ma, and Zhang 2005) and DeepNovo (N. H. Tran et al. 2017). The most popular peptide identification method consists in searching the m/z ratios from the precursor peptides and their fragments against a known database of theoretical fragments. Database search algorithms include Sequest (Eng, McCormack, and Yates 1994), Andromeda (Jürgen Cox et al. 2011) and Protein Prospector (Clauser, Baker, and Burlingame 1999). The database is generated from all protein sequences known to be produced by the genome of the biological system under study. The sequences are first digested *in silico* into peptides, simulating the cleavage rules of the protease used in the experiment (usually trypsin). For each *in silico* peptide, a list of expected fragment masses is then calculated based on the fragmentation technique previously used. The search algorithms calculate match scores between the measured fragmentation spectra and the theoretical database, and the highest-scoring PSMs are taken as candidates for peptide identification. The identifications are then statistically validated often through a target-decoy approach (Elias and Gygi 2007), in which the fragmentation spectra are also searched against random databases for FDR control. One way to generate decoy or random databases is to reverse the target sequences, establishing peptides that do not occur in nature. Finally, the identified peptides are assembled into a list of proteins. However, this process is challenging because some peptides can match multiple proteins (Ahrens et al. 2010). Parsimonious models (Jürgen Cox and Mann 2008; X. Yang et al. 2004; Z. Q. Ma et al. 2009; Slotta, McFarland, and Markey 2010) apply fast heuristics in order to find the smallest set of proteins that explain

the identified peptides. Independently of the chosen method, protein inference accuracy can be improved by applying thresholds on the peptide identification quality, e.g., PSMs with 1% FDR (Sinitcyn, Rudolph, and Cox 2018).

1.6.1.2.2. Bioinformatic analysis of MS data

Once the MS raw data has been pre-processed, the downstream analyses usually involve normalization of the intensities or ratios, filtering of proteins, and imputation of missing values. Clustering analysis, such as hierarchical clustering and k-means, is often used to find groups of proteins with similar abundance patterns along samples or conditions (Ryan et al. 2017; B. Zhang et al. 2014). Principal component analysis (PCA) can be used to explore the relatedness between samples and the existence of confounding effects in the data (Gonçalves et al. 2017). Differentially expressed proteins between two conditions of interest (e.g., tumour vs normal tissues) are often found using a *t*-test or ANOVA (Pozniak et al. 2016; Roumeliotis et al. 2017). However, the DGE method *limma*, addressed in subchapter **1.5.1.4.1**, has also been used in proteomics research (McDermott et al. 2020; Kammers et al. 2015; Schwämmle, León, and Jensen 2013; D'Angelo et al. 2017). Protein co-expression networks can also be constructed and explored using WGCNA (subchapter **1.5.1.4.2**) (H. Zhang et al. 2016; Roumeliotis et al. 2017; Seyfried et al. 2017). When an interesting group of proteins is found, for instance by differential expression or co-expression networks, enrichment analysis can be performed using the methods already addressed in subchapter **1.5.1.4.3**.

1.7. Phosphoproteomics

There are approximately 20,000 protein-coding genes in the human genome (Dunham et al. 2012). However, the corresponding number of human proteins and proteoforms is estimated to be at least one order of magnitude larger than the coding genome would predict (Jensen 2006; Smith and Kelleher 2013). The human proteome is considerably more diverse and complex mainly due to the action of two mechanisms: at the transcriptional level by the alternative RNA splicing (subchapter **1.5**) and after protein translation by covalent PTMs (Walsh, Garneau-Tsodikova, and Gatto 2005). There are more than 200 PTMs characterized to this date, including phosphorylation, acetylation,

ubiquitination and glycosylation (Jensen 2006; Witze et al. 2007). PTMs are so important for the biology of the cell that about 5% of the genes in higher eukaryotes encode enzymes responsible for carrying out the spectrum of PTMs known so far (Walsh, Garneau-Tsodikova, and Gatto 2005).

Protein phosphorylation is the most well-studied PTM (Altelaar, Munoz, and Heck 2013). It is catalysed by protein kinases and involves the reversible transfer of a phosphate group from an ATP molecule to specific residues (phosphorylation site or phosphosite) in the protein substrates, which, in eukaryotes, are usually serine, threonine and tyrosine residues (Ubersax and Ferrell 2007). Serines (Ser or S) are the most frequently modified residues (86.4%), followed by threonines (Thr or T) (11.8%) and tyrosines (Tyr or Y) (1.8%) (Olsen et al. 2006). The opposite reaction (phosphate group removal or dephosphorylation) is catalysed by protein phosphatases (Yusen Liu, Shepherd, and Nelin 2007). Phosphorylation can modulate the function and activity of proteins by altering their 3D conformation, disordered/ordered states and protein-protein binding affinities (Nishi, Shaytan, and Panchenko 2014).

Each kinase is often capable of phosphorylating multiple protein substrates, and, in turn, a given protein can be phosphorylated at multiple positions by the action of different kinases (Nishi, Shaytan, and Panchenko 2014; Miller and Turk 2018). This dynamic interaction between kinases, phosphatases and protein substrates creates an extremely complex protein phosphorylation network that is widespread across signalling transduction pathways (Linding et al. 2008; Newman, Zhang, and Zhu 2014; Mok, Zhu, and Snyder 2011). Phosphorylation is therefore involved in the regulation of most biological processes, including metabolism, cell growth and division, differentiation, organelle trafficking and membrane transport (Ubersax and Ferrell 2007). Not surprisingly, aberrant phosphorylation mediated by uncontrolled kinase/phosphatase signalling has been implicated in many diseases, including cancer (Guha et al. 2008; Deschênes-Simard et al. 2014) and diabetes (Mackenzie and Elliott 2014; Ortsäter et al. 2014). The biological relevance of protein phosphorylation is further demonstrated by the relatively large number of kinase genes (over 500), making up about 2% of the human genome (S. A. Johnson and Hunter 2005; Ubersax and Ferrell 2007).

The phosphoproteome corresponds to the full collection of phosphorylation and dephosphorylation events that dynamically modifies the proteome of a given cell. Phosphoproteomics has the difficult task of cataloging and quantifying such events (Reinders and Sickmann 2005; Hoffert and Knepper 2008). Within the last few years, the development of MS-based phosphoproteomics has been making possible the study of human phosphoproteomes from different cell types and diseases (Ochoa et al. 2020; L. B. Wang et al. 2021), and even phosphoproteomic changes after viral infections (Bouhaddou

et al. 2020). Although the experimental workflows of MS-based phosphoproteomics and proteomics largely overlap, some important considerations have to be made. The following subchapters address them.

1.7.1. MS-based phosphoproteomics

Like in bottom-up proteomics, most MS-based phosphoproteomics approaches use a LC-MS/MS to analyse proteolytic phosphopeptide fragments derived from highly complex protein mixtures (Hoffert and Knepper 2008) (**Figure 1.6**). However, a great difficulty in phosphoproteomics is the low abundance of the modified peptides in comparison to the non-modified peptides (Jürgen Cox and Mann 2011; Choudhary and Mann 2010). This problem can be overcome using diverse strategies that aim to enrich the peptide mixture with phosphorylated peptides (**Figure 1.6C**), such as immobilized metal ion affinity chromatography (IMAC) (Posewitz and Tempst 1999; Villén and Gygi 2008), titanium dioxide (TiO_2) chromatography (Pinkse et al. 2004) and phosphotyrosine immunoprecipitation (Rush et al. 2005). These techniques allow to reduce the complexity of the mixture and detect low abundant peptides, which otherwise may not be detected. After this step, the phosphopeptide mixture is fractionated by an HPLC column, ionized and sprayed into the mass spectrometer, which records the MS^1 and MS^2 spectra of the precursor peptide ions and their fragments, respectively. It is expected that the phosphorylated residues cause a defined m/z change in the peptides, which is directly measured by the mass spectrometer at the MS^1 level. The phosphorylated residues can then be localized with single amino acid resolution at the MS^2 level (Choudhary and Mann 2010).

In order to identify the phosphopeptides, the MS^1 and MS^2 spectra are searched against a protein sequence database. The search for phosphopeptides in sequence databases is nevertheless more challenging than that of non-modified peptides. The size of the search database increases by considering the possibility of phosphorylation at every serine, threonine and tyrosine residues, which can decrease the statistical confidence of the search results (Choudhary and Mann 2010; Jürgen Cox and Mann 2011). Two search approaches were designed to overcome this limitation: open search (Chick et al. 2015) and dependent peptide search (Savitski, Nielsen, and Zubarev 2006). The determination of the exact localization of the phosphorylation site within the protein sequence represents another challenge, particularly when multiple potential phosphosites are close to each other in the sequence. This situation often results in unresolved phosphopeptides with multiple possible

phosphorylation sites. Several software tools provide probabilistic analyses to determine the localization of the phosphosites, including Ascore (Beausoleil et al. 2006), PhosphoScore (Ruttenberg et al. 2008), and SLoMo (C. M. Bailey et al. 2009).

The label-based and label-free quantification methods used in proteomics allow to determine the absolute abundance of a given phosphosite in a biological sample or its relative change between two or more conditions (see subchapter **1.6.1.1**). Phosphosite quantification is nevertheless less accurate than protein quantification because of the lower number of features (phosphopeptides) usually available for quantification, which increases the number of missing values and the variance of the data across the cohorts (Sinitcyn, Rudolph, and Cox 2018).

An important concept that must be taken into account in phosphoproteomics is the stoichiometry of a given phosphosite, which is defined as the fraction of total protein that is phosphorylated at that site. The stoichiometry of a phosphorylation site, in a given condition, may help in understanding its functional relevance in that condition. For instance, phosphosites associated with protein function are expected to have stoichiometries correlated with the activity of the corresponding proteins. Phosphosites that regulate a larger fraction of total protein will therefore be involved in regulating to a larger extent the protein activity, protein localization and protein-protein interactions (Prus et al. 2019). Phosphosite abundance can be confounded by the respective protein levels, i.e., changes in the abundance of a phosphosite can be driven by variation in the total protein abundance and not by variation in the site stoichiometry (Altelaar, Munoz, and Heck 2013). To counteract this problem parallel protein quantification is often required, which can then be used to regress-out the protein abundance from the phosphorylation data. Regressing-out the protein abundance involves calculating the residuals of a linear model fitted between the phosphosite and the protein quantifications, as dependent and independent variables, respectively (Roumeliotis et al. 2017).

Phosphorylation quantification data can be analysed with bioinformatic methods similar to those used for protein quantification (see subchapter **1.6.1.2.2**). As noted previously, the phosphorylation-level data has two disadvantages compared to protein-level data, which are the high variance and number of missing values. The high variance can be taken into account beforehand by using a higher number of replicates in order to increase the statistical power (Sinitcyn, Rudolph, and Cox 2018). The problem of data sparsity can be circumvented by avoiding analysing individual phosphosites and focusing instead on predicting the activities of kinases, as pointed out by (Ochoa et al. 2016). Methods for kinase activity prediction are discussed in the following subchapter.

1.7.2. Kinase activity prediction

Each kinase binds to specific phosphorylation sites on their protein substrates, where they catalyse the phosphorylation reactions. During kinase-substrate binding, kinases must recognize a few hundred compatible phosphorylation sites within a background of hundreds of thousands to millions of phosphorylatable residues. The specificity of kinases for certain substrates in detriment of others is driven by several mechanisms, including the depth of the kinase catalytic cleft, which determines the specificity for serine, threonine or tyrosine residues; the affinity for certain consensus sequences around the phosphosite; docking with additional motifs on the substrate; and interaction with protein adaptors or scaffolds (Ubersax and Ferrell 2007).

Several databases have been created to store experimentally validated phosphorylation sites and kinase-substrate interactions derived from *in vitro* and *in vivo* studies. These databases include PhosphoSitePlus (Hornbeck et al. 2015) and Signor (Perfetto et al. 2016). As of March 2021, PhosphoSitePlus stores 12,841 manually-annotated kinase-substrate human interactions involving 401 kinases. The Signor database contains 9,747 interactions for 460 kinases. Apart from reporting the phosphorylation reactions, Signor also indicates whether they activate or inhibit the substrates. Other resources, such as ProtMapper (Bachman, Gyori, and Sorger 2019) and OmniPath (Türei, Korcsmáros, and Saez-Rodriguez 2016), were developed to aggregate interactions from multiple sources. ProtMapper compiles interactions not only from the aforementioned databases but also from HPRD (Mishra et al. 2006), BEL Large Corpus (https://bel.bio/), NCI-PID (Schaefer et al. 2009) and Reactome (Croft et al. 2014), and from the text-mining tools REACH (Valenzuela-Escárcega et al. 2018), RLIMS-P (Torii et al. 2015) and Sparser. Omnipath combines together data from more than 100 resources. The data consists in protein-protein interactions and targets of kinases, phosphatases, TFs and drugs.

The activity of a kinase is reflected in the phosphorylation status of the substrates that it targets. The integrated analysis of MS-based phosphosite quantifications with prior knowledge about kinase-substrate interactions has been used to computationally estimate the activity of kinases. Different methods have been developed to estimate the activity of kinases based on the change of phosphorylation levels of their substrate sites. In a benchmark study performed by (Hernandez-Armenta et al. 2017), it was found that the kinase set enrichment analysis (KSEA) and the one sample Z-test have the best overall performance. These methods were successfully applied to estimate the kinase activity profiles of cancer cells (Casado et al. 2013; Drake et al. 2012) and to broadly survey the kinase signalling states of human cells (Ochoa et al. 2016).

The KSEA algorithm is similar to that used for GSEA (see subchapter **1.5.1.4.3**). KSEA considers a ranked list of phosphosite quantifications in order to assess whether a predefined set of kinase substrates is statistically enriched in phosphosites that are at the two extremes of the list (Ochoa et al. 2016; Drake et al. 2012). KSEA runs as follows: (i) an enrichment score (ES) is calculated by walking down the ranked phosphosite list and updating a weighted Kolmogorov-Smirnov-like statistic. The statistic is increased when it encounters a substrate of the kinase and vice-versa; (ii) the null distribution of the ES is computed by randomizing the phosphosite labels (e.g., 10,000 times) and re-calculating the ES; (iii) an empirical P-value for the observed ES is calculated from the null distribution.

The Z-test compares the mean quantification (e.g., fold-changes) of the substrates of a given kinase to the mean and variance of all phosphosites quantified in a given sample (Hernandez-Armenta et al. 2017; Casado et al. 2013; S. Y. Kim and Volsky 2005). The Z-test is calculated as follows:

Equation 1.5: $z = \frac{x - \mu}{\sigma / \sqrt{N}}$

where z corresponds to a z-score, x the mean quantification of the kinase substrates, μ and σ the mean and standard deviation of all phosphosites quantified in the sample, respectively, and $\sqrt{N}$ the square root of the number of kinase substrates. A P-value is then calculated from the z-score using the standard normal distribution.

These tests generate P-values that reflect the statistical significance of kinase regulation. To obtain a kinase activity score that indicates whether the kinase is increasing or decreasing activity, the P-values are -log10 transformed and signed based on the mean fold-change of the kinase substrates in order to account for kinase activation or deactivation.

1.8. Aims of the thesis

As previously discussed in subchapter **1.2**, the TCGA and CPTAC consortia have provided crucial results that shed light on how DNA alterations contribute to the onset and progression of cancer. Both projects have harnessed the power of *omics* to build state-of-the-art cancer maps describing in detail the numerous ways that genetic alterations lead to dysregulated cellular pathways, thereby altering the controlled growth and division of healthy cells and rendering them cancerous. Equally important, these projects have made available to the research community an enormous amount of multi-omics datasets that

broadly cover multiple layers of biological information, including the genome, transcriptome and phospho(proteome) of cancer cells. This thesis aims to take advantage of the multi-omics data generated in the scope of TCGA and CPTAC in order to study the heterogeneity of gene expression, protein abundance and protein activities across multiple cancer types. It is expected that this new knowledge will be translated into improved clinical care and monitoring of cancer patients. The aims of this thesis are:

- Characterize the gender differential transcriptome in stomach and thyroid cancer, two cancer types with a clear unbalanced gender incidence. The results of this work are described in chapter **2** and include published material from the following article: *Abel Sousa, Marta Ferreira, Carla Oliveira and Pedro G. Ferreira. Gender Differential Transcriptome in Gastric and Thyroid Cancers. Frontiers in Genetics, Volume 11, Article 808, 30 July 2020.*

- Describe the protein-level attenuation of copy-number alterations, combining genomics, transcriptomics and (phospho)proteomics cancer data. These results are described in chapter **3** and were published in the following article: *Abel Sousa, Emanuel Gonçalves, Bogdan Mirauta, David Ochoa, Oliver Stegle and Pedro Beltrão. Multi-omics Characterization of Interaction-mediated Control of Human Protein Abundance levels. Molecular & Cellular Proteomics, Volume 18, Issue 8, Pages 114-125, 9 August 2019.*

- Analyse the impact of genetic alterations on the activities of kinases and TFs. These results are reported in chapter **4** and are available in the following article: *Abel Sousa, Aurelien Dugourd, Danish Memon, Borgthor Petursson, Evangelia Petsalaki, Julio Saez-Rodriguez and Pedro Beltrão. Pan-Cancer Landscape of Protein Activities Identifies Drivers of Signalling Dysregulation and Patient Survival. BioRxiv, 9 June 2021.*

Finally, chapter **5** is dedicated to discuss and conclude the research performed in this thesis, including potential future directions to continue studying the aforementioned biological subjects using multi-omics cancer datasets.

2. Gender Differential Transcriptome in Gastric and Thyroid Cancers

This chapter includes published material from the following article:
Abel Sousa, Marta Ferreira, Carla Oliveira and Pedro G. Ferreira. Gender Differential Transcriptome in Gastric and Thyroid Cancers. Frontiers in Genetics, Volume 11, Article 808, 30 July 2020.

2.1. Abstract

Cancer has an important and considerable gender differential susceptibility confirmed by several epidemiological studies. Gastric (GC) and thyroid cancer (TC) are examples of malignancies with higher incidence in males and females, respectively. Beyond environmental predisposing factors it is expected that gender-specific gene deregulation contributes to this differential incidence. We performed a detailed characterization of the transcriptomic differences between genders in normal and tumour tissues from stomach and thyroid, using Genotype-Tissue Expression (GTEx) and The Cancer Genome Atlas (TCGA) data. We found hundreds of sex-biased genes (SBGs). Most of the SBGs shared by normal and tumour belong to sexual chromosomes, while the normal and tumour-specific tend to be found in the autosomes. Expression of several cancer-associated genes is also found to differ between sexes in both types of tissue. Thousands of differentially expressed genes (DEGs) between paired tumour-normal tissues were identified in GC and TC. For both cancers, in the most susceptible gender the DEGs were mostly under-expressed in the tumour tissue, with an enrichment for tumour suppressor genes (TSGs). Moreover, we found gene networks preferentially associated to males in GC and to females in TC and correlated with cancer histological subtypes. Our results shed light on the molecular differences and commonalities between genders and provide novel insights in the differential risk underlying these cancers.

2.2. Introduction

Sexual dimorphism is a taxonomically widespread phenomenon, whereby certain traits differ consistently between males and females within a given species. In humans and other animals, these differences go beyond morphological and behavioural traits and include molecular phenotypes such as gene expression (Trabzuni et al. 2013; Melé et al. 2015; Gershoni and Pietrokovski 2017; Naqvi et al. 2019). It has been hypothesized that sex-specific gene regulation underlies important phenotypic gender differences and may contribute to gender-differential susceptibility to disease (Ober, Loisel, and Gilad 2008; Rawlik, Canela-Xandri, and Tenesa 2016; Labonté et al. 2017). Cancer has a considerable differential incidence between genders (Tevfik Dorak and Karpuzoglu 2012; Clocchiatti et al. 2016; Ali et al. 2016), with men showing higher cancer incidence than women in 32 of 35 anatomical sites (Edgren et al. 2012). In 13 of these sites, the differences could not be

explained by known risk factors, including smoking, alcohol consumption and potential occupational carcinogens such as toxic metals and ionizing radiation. Men are at higher risk and worst prognosis in several types of cancers in non-reproductive tissues, including skin, esophagus, stomach, liver and urinary bladder cancers (Siegel, Miller, and Jemal 2018). One remarkable exception is the thyroid tissue, where women have three times higher risk of developing cancer (Rahbari, Zhang, and Kebebew 2010). For malignancies such as acute lymphoblastic leukemia or non-Hodgkin lymphoma, the gender-bias incidence occurs already in childhood, being more common in boys (Tevfik Dorak and Karpuzoglu 2012). Although environmental and lifestyle factors largely contribute to gender disparities in cancer, it seems clear that gender intrinsic molecular factors may also play an important role.

Cancer sexual disparity may be the consequence of a complex interplay between sex chromosomes and the hormonal system (Clocchiatti et al. 2016). In females, several X chromosome genes may escape the XIST-dependent inactivation, triggering an imbalanced expression between genders (Carrel and Willard 2005). This asymmetry can make females more resistant to inactivating mutations in tumour-suppressor genes (TSGs) (Dunford et al. 2017). For instance, *UTX* is known to escape silencing in females (Bellott et al. 2014) and to have inactivating mutations in renal and esophageal cancers, more prevalent in males (Van Haaften et al. 2009). Sex steroid hormones can interact with the cellular receptors estrogen receptor-α (ERα), ERβ and androgen receptor (AR), and induce gene expression changes, affecting cellular metabolic states, tumour microenvironments and the immune system (Clocchiatti et al. 2016). For example, in liver cancer, more frequent in males, AR stimulates and ERα restrains cellular proliferation (Z. Li et al. 2012). Moreover, an estrogen-mediated inhibition of inflammatory IL-6 production may reduce liver cancer risk in females (Naugler et al. 2007). In thyroid cancer (TC), the association between sex hormones and cancer risk is uncertain (Rahbari, Zhang, and Kebebew 2010). While animal models and *in vitro* studies suggest that sex hormone levels can affect TC tumorigenesis and progression, the same has not been observed at the clinical level (Yao et al. 2011). Sex hormones are also known to regulate the thyroid gland in a gender-specific manner (Banu, Govindarajulu, and Aruldhas 2002). It is therefore possible that the thyroid glands in females are biologically more prone to cancer development than in males (Yao et al. 2011).

Pan-cancer systematic studies on gender differences have identified sex-biased genes and pathways across several cancer types from The Cancer Genome Atlas project (TCGA) (J. Ma, Malladi, and Beck 2016; Yuan et al. 2016). These studies found that sex-specific gene signatures have differential responses to chemical and genetic agents, and that, in certain cancer types, more than 50% of clinically relevant genes are differentially

expressed between sexes. Importantly, while TC showed extensive sex-biased gene expression, the gender differences of gastric cancer (GC) remained uncharacterized.

In this work, we set out to provide a fine-detailed characterization of the gender differential transcriptome in GC and TC (Bass et al. 2014; Agrawal et al. 2014), chosen due to their clear unbalanced gender incidence. While GC is two times more common in males, TC is three times more common in females (Siegel, Miller, and Jemal 2018). Our results demonstrate that sex-biased gene expression is more pronounced in normal tissues than tumour tissues, and that most of the shared variation arises from the sexual chromosomes. Expression of several cancer-associated genes differs between genders, with TSGs preferentially down-regulated in the tumour tissue of the most susceptible gender. Gene co-expression network analysis revealed an extensive topological preservation between genders, with gender-specific networks appearing correlated with cancer histological subtypes.

2.3. Results

2.3.1. Gender differences are not revealed by genome-wide transcriptomic profiles

We analysed RNA-seq data, generated by the TCGA project, of 375 GC samples (female, n=134 and male, n=241) and 502 TC samples (female, n=367 and male, n=135). As normal tissue counterparts, we used data from the Genotype-Tissue Expression project (GTEx) V6 (GTEx Consortium, 2013), that encloses 225 normal stomach samples (female, n=82 and male, n=111) and 381 normal thyroid samples (female, n=112 and male, n=211), as well as the TCGA tumor-matched normal samples (stomach, n=32 and thyroid, n=58). Altogether, we collected 1,483 tumour and normal samples from stomach and thyroid tissues (**Supplementary figure 2.1A; methods**). Minimal expression filtering yielded 11,842 genes in stomach and 11,734 in thyroid (**Methods**).

Principal component analysis (PCA) analysis revealed that both tissues segregated by dataset of origin rather than tumour or normal status (**Supplementary figure 2.1B**). As a consequence of this strong batch effect, all analyses have been performed separately for the TCGA and the GTEx datasets. No global distinct transcriptomic patterns were observed

between genders (**Supplementary figure 2.1B**). Confounding effects were successfully regressed-out (**Supplementary figure 2.1C, Supplementary figure 2.2**).

A detailed characterization of the transcriptomic differences between genders was performed following the design in **Figure 2.1**.

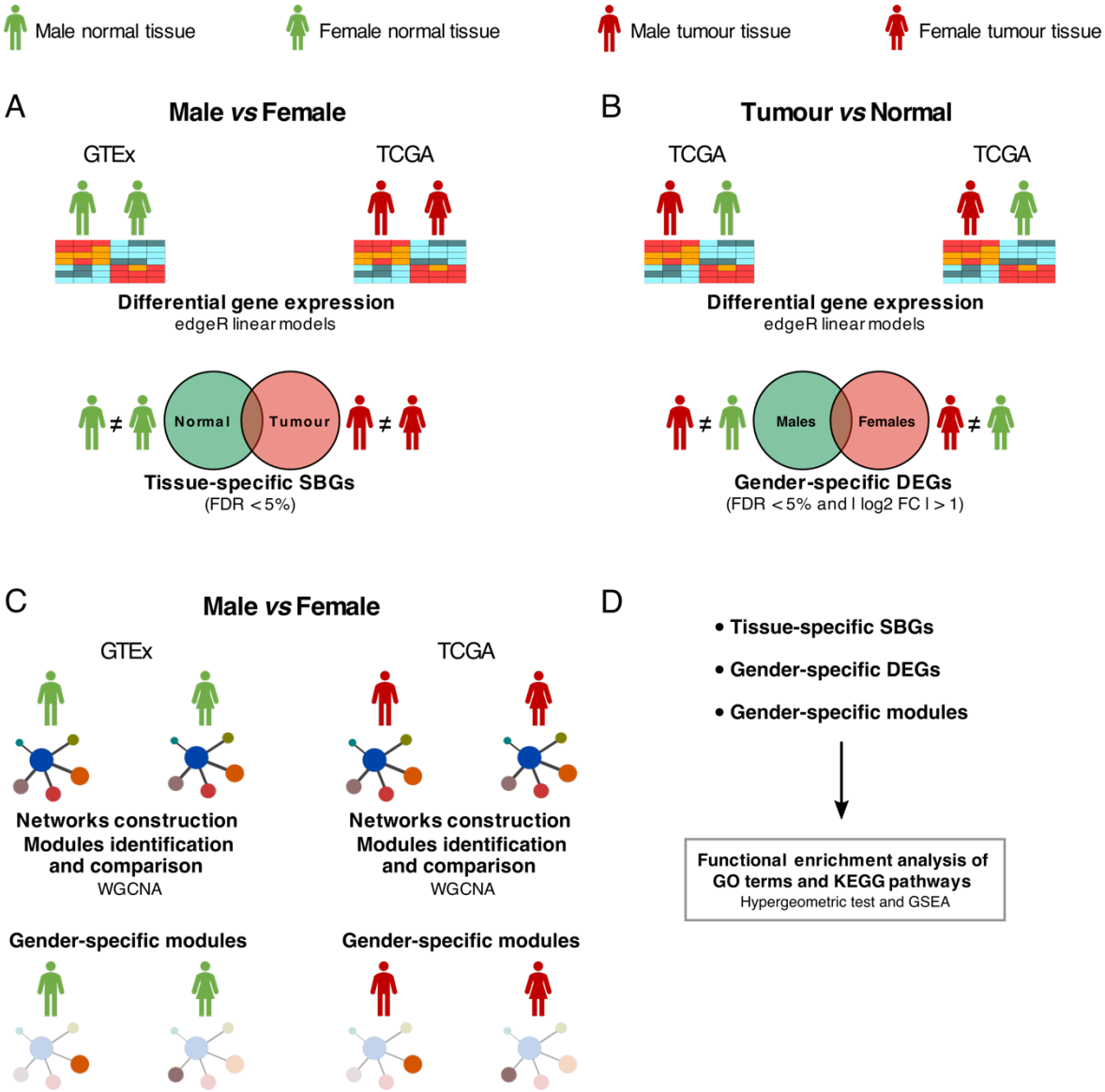


Figure 2.1. Study design. (A) Differential expression analysis between males and females in normal (GTEx) and tumour (TCGA) samples, adjusted for confounding effects. SBGs: sex-biased genes. **(B)** Differential expression analysis between tumour and matched-normal (TCGA) samples in males and females, adjusted for confounding effects. DEGs: differentially expressed genes. **(C)** Differential co-expression network analysis between males and females in normal (GTEx) and tumour (TCGA) samples. WGCNA: Weighted Correlation Network Analysis. **(D)** Functional enrichment analysis of SBGs, DEGs and gene co-expression modules was performed using hypergeometric-based tests and Gene Set Enrichment analysis (GSEA).

2.3.2. Tumour and normal tissues show specific sex-biased genes

To understand the gender and tissue-specific (tumour and normal) expression patterns in stomach and thyroid, we performed gender-differential expression analysis using the normal samples from both tissues available from GTEx, followed by the same analysis in the tumour samples from TCGA (**Figure 2.1A**, **Figure 2.1D**). In stomach we found 75 sex-biased genes (SBGs) in the normal and 55 SBGs in the tumour, of which 32 were common (**Figure 2.2A, Figure 2.2C; Table S2.1**). For thyroid we found 691 and 128 SBGs in the normal and tumour, respectively, with 46 genes in common (**Figure 2.2B, Figure 2.2E; Table S2.2**). Common SBGs originated mostly from the X and Y chromosomes (**Figure 2.2D, Figure 2.2F**) and were similar in stomach and thyroid (27 genes; 84% and 59% of the common SBGs in stomach and thyroid). These genes were involved in translational initiation, protein dealkylation and demethylation, with preferential location in the genomic regions chrXp22/p11/q13 and chrYq11 (**Supplementary figure 2.3**; false discovery rate [*FDR*] < *5%*). Contrarily, normal and tumour-specific gender differences derived mostly from autosomes (**Figure 2.2D, Figure 2.2F**).

The 43 normal-specific SBGs in stomach (**Figure 2.2C**) were enriched for sterols metabolic processes and in the peroxisome proliferator-activated receptor (PPAR) signalling pathway (**Figure 2.2G;** *FDR* < *5%*). The majority of genes involved in these processes were over-expressed in the normal stomach of females, the less affected gender in GC (**Supplementary figure 2.4A**). In thyroid the 645 normal-specific SBGs (**Figure 2.2E**) were enriched in innate and adaptive immune response processes and lipids metabolism (**Figure 2.2H, Supplementary figure 2.5A;** *FDR* < *5%*), including over-expression in females, the most affected gender in TC (**Supplementary figure 2.4B, Supplementary figure 2.5B**).

The 23 tumour-specific SBGs in GC and 82 in TC (**Figure 2.2C, Figure 2.2E**) were involved in lipids metabolic processes (**Supplementary figure 2.6;** *FDR* < *20%*).

The X-chromosomal SBGs from GC and TC showed an enrichment for genes that escape X-inactivation (**Table S2.3;** *P-value* < *5%*). Of note, *USP9X (log2FC/FDR = -0.31/1.7e-06)*, a previously reported cancer driver, *TXLNG (-0.54/3.3e-17), OFD1, MED14 and CDK16,* are known to evade X-inactivation and were over-expressed in females' GC. This suggests that tumorigenesis in females' stomach may take advantage from over-expression of genes that escape X-inactivation.

Gender-differential promoter methylation analysis showed that 54% (GC) and 20% (TC) of the previously found SBGs were differentially methylated (**Supplementary figure 2.7A, Supplementary figure 2.7B; Methods;** *P-value = 2e-15, 2.7e-6*). Among these, 96% belong to the X-chromosome and 89% are known to escape X-inactivation.

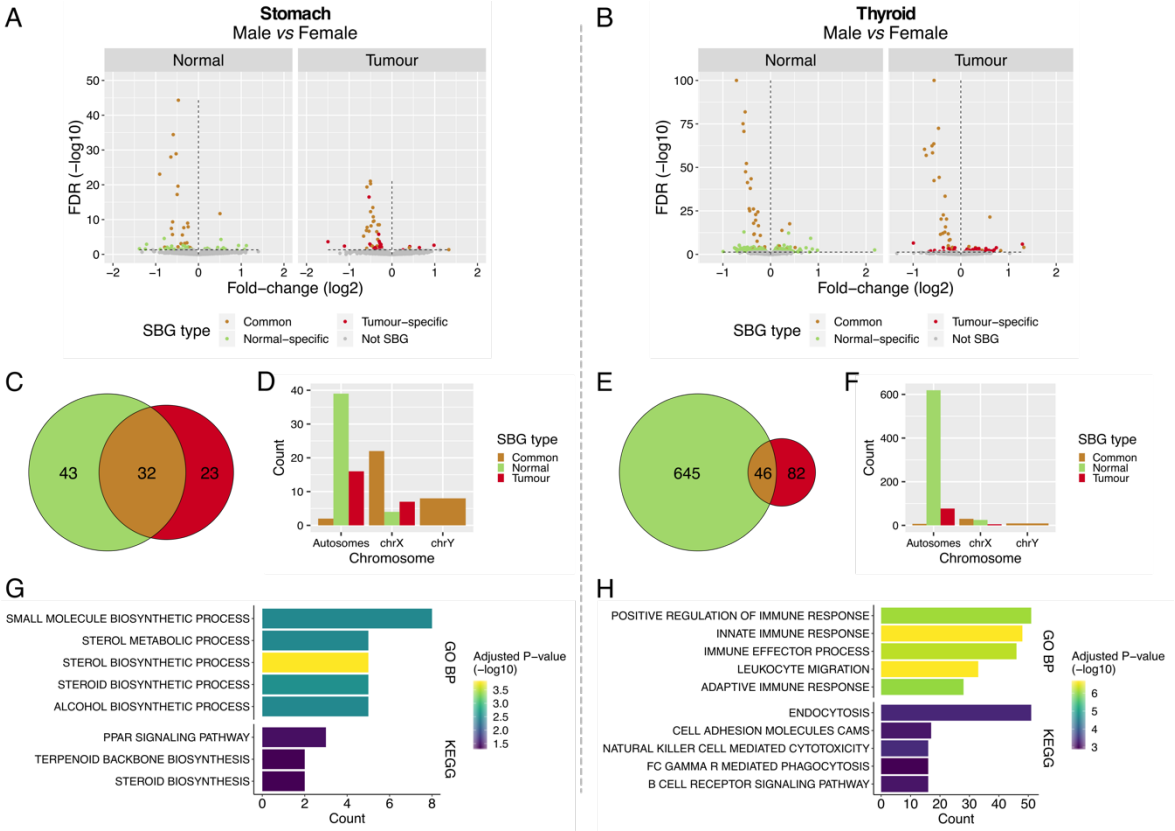


Figure 2.2. Features of sex-biased genes (SBGs) in tumour and normal stomach and thyroid tissues. (A) (B) Differentially expressed genes between males and females, in normal (GTEx) and tumour (TCGA) tissues from stomach and thyroid. The Y-chromosome genes and XIST were removed for visualization purposes. Vertical lines: left - genes over-expressed in females (under-expressed in males); right - genes under-expressed in females (over-expressed in males). **(C) (E)** Shared and tissue-specific SBGs. **(D) (F)** Distribution of SBGs in autosomes and sexual chromosomes. **(G) (H)** Gene Ontology (GO) biological processes (GO BP) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched in the normal-specific SBGs (top 5; FDR < 5%).

2.3.3. Tumour suppressor genes show tumour-specific under-expression in the susceptible gender

To identify tumour-normal differentially expressed genes (DEGs) in each gender, tumour and matched-normal TCGA samples were compared (**Figure 2.1B**, **Figure 2.1D; Tables S2.4**, **S2.5**). In GC we found 1552 DEGs shared between genders, corresponding to 84% of the female and 68% of the male DEGs (**Figure 2.3A, Figure 2.3C**). Similarly, in TC 89% of the female and 68% of the male DEGs were common to both genders (1023 DEGs) (**Figure 2.3B, Figure 2.3E**). The shared DEGs likely reflect genes that, independently of the gender, are pivotal for tumorigenesis. In fact, a significant proportion

of these are oncogenes (5% for GC and 6% for TC; *P-value = 0.02, 9e-3*). In GC it includes the cancer drivers *WHSC1, CBFB, RUNX1, EZH2 (male, female log2FC/FDR = 1.6/2.1e-10, 1.9/6e-7), MET* and *CARD11*. In TC we recapitulated *MET (2.6/2.2e-10, 2.5/3.1e-25)* and *RUNX1*, plus *CCND1, CDKN1A, ERBB3, FOXQ1, FGFR3* and the known oncogene in TC *ZCCHC12* (O. Wang et al. 2017).

Gender-specific DEGs were frequently under-expressed in tumours of the most susceptible sex (males in GC and females in TC), but not in the less susceptible one (**Figure 2.3D, Figure 2.3F** [left plots]). Of 742 male-specific DEGs in GC and 125 female-specific DEGs in TC, 64% and 70% were under-expressed in tumours, respectively. Following this trend, we found a significant enrichment for TSGs on males in GC (9%) and females in TC (17%) (**Figure 2.3D, Figure 2.3F** [left plots]), with the majority being under-expressed in tumours (**Figure 2.3D, Figure 2.3F** [right plots]).

Overall, our results showed that for these two cancers the majority of tumour-normal DEGs were shared by genders and have oncogenic properties. Most of the gender-specific DEGs were under-expressed in tumours from the most susceptible gender with a significant fraction being TSGs.

Promoter methylation analysis between tumour and matched-normal tissues of TC (data not available for GC) showed that in females and males, 53% and 26% of the DEGs were differentially methylated (**Supplementary figure 2.7C, Supplementary figure 2.7D; Methods;** *P-value = 1e-3, 1.7e-24*).

2.3.4. Tumour-normal DEGs were enriched for functional gene categories

Functional enrichment analysis showed that in GC, gender-common DEGs were involved in muscle structure development and contraction, female-specific in cellular responses to cytokine stimulus and male-specific in epithelial cell differentiation and metabolic processes (**Supplementary figure 2.8A;** *FDR < 5%*). In TC, gender-common DEGs were involved in positive regulation of cellular proliferation and pathways in cancer, female-specific in regulation of cell adhesion and T-cell receptor signalling pathways, and male-specific in response to cytokines and innate immune response processes (**Supplementary figure 2.8B;** *FDR < 5%*).

In TC female-specific DEGs over-expressed in normal tissues and male-specific DEGs over-expressed in tumour tissues were involved in similar processes and pathways (**Figure 2.3H;** *FDR < 5%*). Male-specific DEGs over-expressed in normal tissues were

enriched for ion transmembrane transport activity (**Figure 2.3H;** *FDR < 5%*). In GC there were no clear patterns, with gender-specific DEGs showing distinct and diverse functions (**Figure 2.3G**).

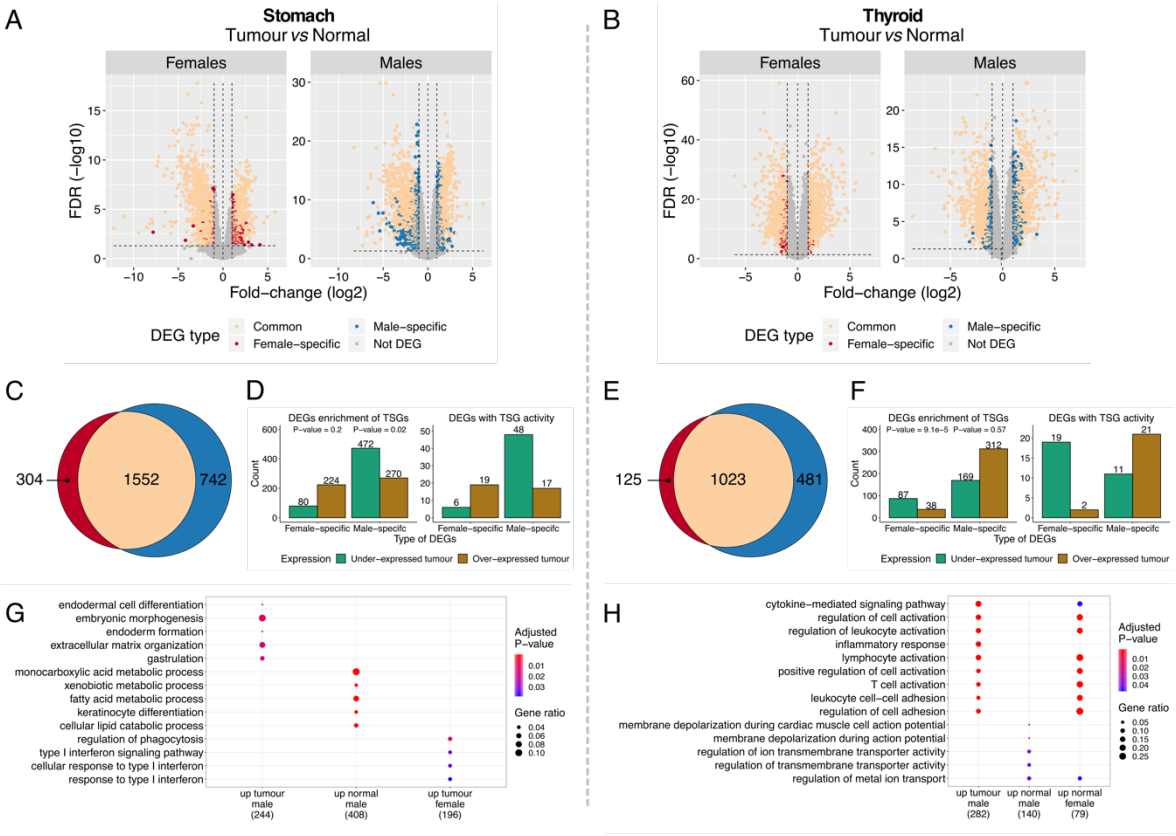


Figure 2.3. Features of differentially expressed genes (DEGs) between tumour and normal tissues in GC and TC. (A) (B) DEGs between tumour and matched-normal samples, in females and males. Vertical line (x = 0): left - genes under-expressed in tumours (over-expressed in normal); right - genes over-expressed in tumours (under-expressed in normal). **(C) (E)** Shared and gender-specific DEGs. **(D) (F)** Gender-specific DEGs over- and under-expressed in tumours. Left: all DEGs and respective enrichment for TSGs (Fisher-test P-value). Right: DEGs with TSG activity. **(G) (H)** GO biological processes (GO BP) significantly enriched and shared between gender-specific DEGs (male/female) over-expressed in tumour and normal tissues (up) (FDR < 5%).

2.3.5. Gender-specific gene networks in cancer are associated with histological subtypes

Co-expression network analysis identifies groups of genes, called network modules, coherently expressed across samples. Such modules may highlight biologically-related genes. We reasoned that beyond single gene sex-biased expression, there are differences between genders regarding the coordinated expression of groups of genes, for tumour and

normal tissues (**Figure 2.1C, Figure 2.1D**). After removing possible confounding effects (**Supplementary figure 2.1C, Supplementary figure 2.2**), a full gene co-expression network was built and modules identified for each gender (**Supplementary figure 2.9, Supplementary figure 2.10; Methods**). For stomach we found 23 modules in normal tissue (Female: 14; Male: 9) and 56 in tumours (F: 22; M: 34) (**Supplementary figure 2.11A**). In thyroid we found 75 modules in normal tissue (F: 37; M: 38) and 39 in tumours (F: 18; M: 21) (**Supplementary figure 2.11A**). The number of genes inside modules ranged from 21 to 5976, with a median size of 119 genes per module (**Supplementary figure 2.11B**). Next, modules were compared between genders in terms of their overlap and deemed as preserved (lowly, moderately or highly) or gender-specific (**Supplementary figure 2.12; Methods**). Most modules were preserved in tumour and normal tissues (**Figure 2.4A**, **Figure 2.4F**). Three female-specific modules were found in normal thyroid, related to vasculature development and angiogenesis; thyroid hormone and sterol metabolism (**Supplementary figure 2.13**).

Consistent with the higher sex-biased cancer incidence, we found 3 male-specific modules in GC and 2 female lowly-preserved modules (in males) in TC (**Figure 2.4A**, **Figure 2.4F**). Correlation of the modules representative expression profile with the cancer clinical traits (**Methods**), revealed one GC module (M2S, *P-value* = *4e-3*) and one TC module (M2T, *4e-18*) associated with the cancer histological subtypes (**Figure 2.4B**, **Figure 2.4G**). The former (M2S) involved genes related to peptide metabolism and translation elongation (**Figure 2.4C;** *FDR* < *5%*). The latter (M2T) was related to cellular respiration processes**,** with the most highly connected (hub) genes forming part of the oxidative phosphorylation pathway (**Figure 2.4H;** *FDR* < *5%*). A higher intra-module correlation was found for the gender where the module is specific (**Figure 2.4D**, **Figure 2.4I;** *P-value* < *2.2e-16*), reflecting considerably different network topologies (**Figure 2.4E**, **Figure 2.4J**). These results demonstrate that the coordinated expression of these genes differ between genders.

Hub genes from gender-specific modules were associated with specific cancer histological subtypes. In GC, 11 out of 12 hub genes showed over-expression in the papillary, tubular and non-specified intestinal subtypes of males (**Supplementary figure 2.14A, Supplementary figure 2.14C; Table S2.6)**. Among these, *Hsp90ab1* and *XPO5* have been previously associated with poor prognosis and tumour-suppressor properties in GC (H. Wang et al. 2019; Melo et al. 2010). In TC, all 27 hub genes were predominantly over-expressed in the follicular subtype of females (**Supplementary figure 2.14B, Supplementary figure 2.14D; Table S2.7**).

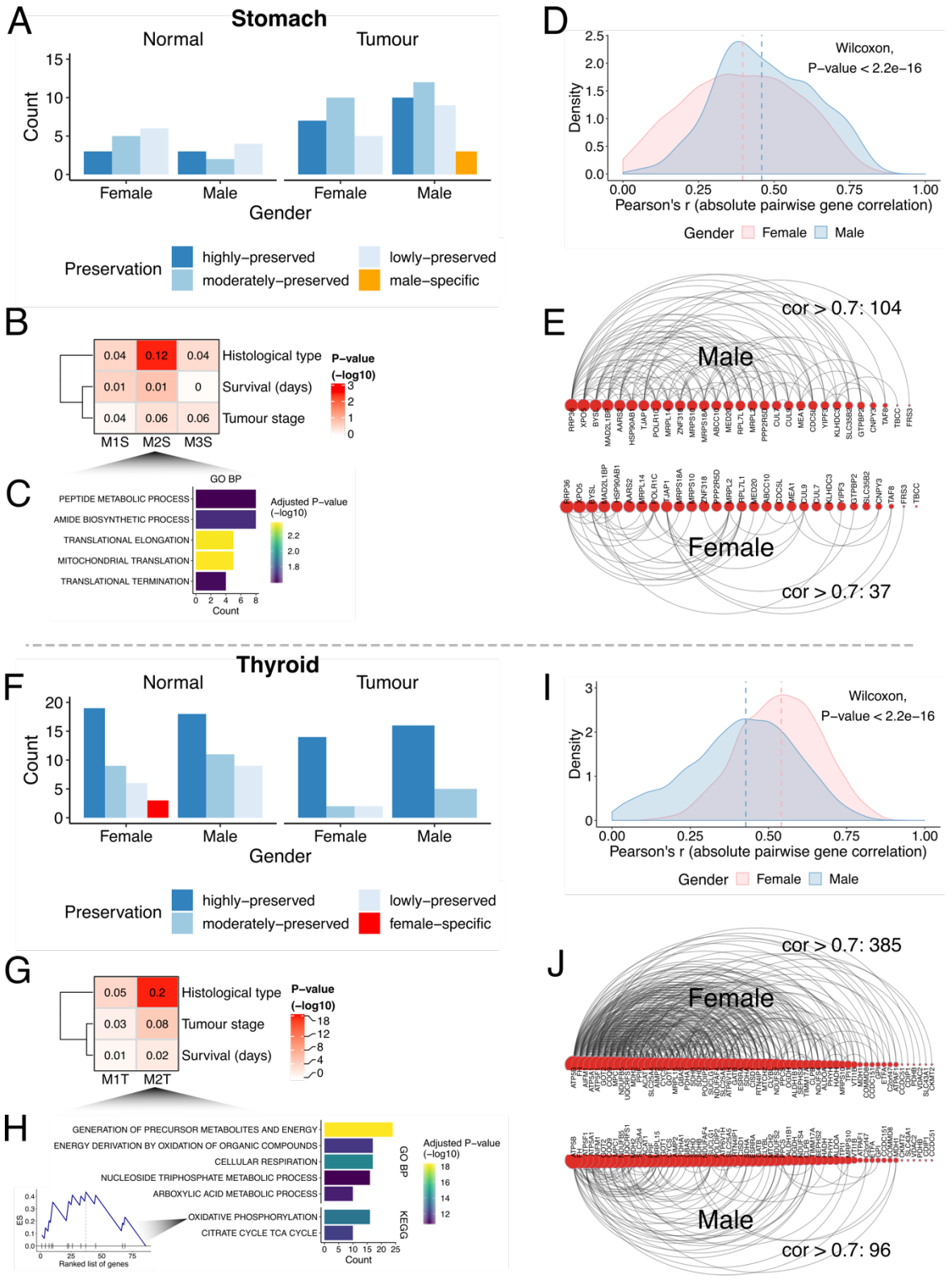


Figure 2.4. Gender-differential co-expression network analysis. (A) (F) Number of modules in stomach and thyroid tissues. **(B) (G)** Association between the GC male-specific modules and the TC female lowly-preserved (in males) modules with cancer clinical traits. The numbers inside the heatmaps are regression-derived R^2. The one-way ANOVA-derived P-value is shown (-log10). Associations with survival were also tested using cox hazard regressions (log-rank P-values > 5%). **(C) (H)** GO biological processes (GO BP) and KEGG pathways enriched in the GC M2S module and the TC M2T module (top 5; FDR < 5%). Enrichment score (ES) for the oxidative phosphorylation pathway in M2T is highlighted, with

genes sorted by intra-modular connectivity degree. **(D) (I)** Distribution of the pairwise gene-correlations (absolute Pearson's r) for all gene pairs in M2S and M2T. Vertical lines indicate medians. **(E) (J)** Arc diagrams representing gene-pair correlations for M2S (29 genes from 45) and M2T (61 genes from 86). Arcs represent gene-pair correlations > 0.7. Genes are sorted by number of connections. The gene-pairs with correlations > 0.7 were selected in the gender where the module shows specificity (males in M2S and females in M2T). The number of correlations > 0.7 decreases in the opposite gender.

2.4. Discussion

In this work, we set out to characterize the gender differential transcriptome in tumour and normal tissues. We selected GC and TC due to their considerable gender-biased incidence. It is well known that male and females are exposed often to very different environmental conditions (Scarselli et al. 2018; Zahm and Blair 2003). Thus, an important limitation of this study is the lack of control for environmental effects. Despite this, we believe this analysis is still of merit since it may capture intrinsic natural variation between genders and their relation to disease susceptibility.

Our results show that SBGs in tumour and normal tissues were mostly derived from sex chromosomes, as previously found in (Yuan et al. 2016). These genes are mostly common to stomach and thyroid. Such conservation remains to be tested for other tissues. On the other hand, SBGs specific to normal or tumour tissues arose from non-sexual chromosomes, with little overlap between tissues. These results highlight the contribution of autosomes for tumour- and normal-specific sex-biased expression phenotypes, which may ultimately drive sex-biased cancer incidence.

We found metabolic processes of sterols and lipids enriched in the normal- and tumour-specific SBGs of thyroid and stomach, and in a female-specific module of thyroid normal tissues. Sterols are critical in signalling, regulation of lipids metabolism, development and cellular homeostasis (Wollam and Antebi 2011). Alterations in the metabolism of sterols and lipids is a known hallmark of cancer (Gabitova, Gorin, and Astsaturov 2014). Other studies have found sex-biased genes in pathways related to the metabolism of fatty acids, in cancer (Yuan et al. 2016) and in normal tissues (Naqvi et al. 2019), with a long-standing observation that genders show differences in the metabolism of lipids (Mittendorfer 2005; Mittelstrass et al. 2011; Drolz et al. 2014). Importantly, such differences are not simply explained by the presence and action of sex hormones (Mittendorfer 2005). Our results suggest that beyond differences in sexual hormonal

regulation, the metabolic physiology of the sexes might be implicated in the gender disparity of GC and TC.

The PPAR signalling pathway was found enriched in the normal-specific SBGs of stomach and particularly over-expressed in females. This pathway is known to control the expression of genes involved in lipids metabolism and inflammation (Varga, Czimmerer, and Nagy 2011), with increasing evidence that PPARα/γ inhibits tumour progression and acts as tumour suppressor (Gou et al. 2017). Whether this finding is related to the lower GC incidence in females remains to be seen.

The normal-specific SBGs of thyroid were enriched in immune-response pathways and mostly over-expressed in females, in accordance with (Naqvi et al. 2019). Thyroid hormones can trigger different responses in diverse immune cells and affect several inflammation-related processes (Jara et al. 2017). The immune system is a highly sexually dimorphic trait, with females showing immunological advantage when facing different immune challenges (Libert, Dejager, and Pinheiro 2010). On the other hand, females are more prone to autoimmune diseases such as Hashimoto's thyroiditis (HT) (Ngo, Steyn, and McCombe 2014). In the last decades the association between HT and TC has been growing, with some studies reporting the co-existence of both diseases (Felicetti, Catalano, and Fortunati 2017; L. Zhang et al. 2012; Jeong et al. 2012).

Altered expression of oncogenes and TSGs in normal tissue may be linked to protective or predisposing tumorigenic events (Muir and Nunney 2015). In the stomach, females over-expressed the TSGs *FGFR3* and *ERCC2* (Lafitte et al. 2013; J. F. Zheng et al. 2015). In the thyroid, 16 of the SBGs were previously reported as cancer drivers, with 11 being over-expressed in females' normal thyroid, including the oncogenes *CARD11*, *EZH2* and *IL7R* (Watt et al. 2015; K. H. Kim and Roberts 2016; M. J. Kim et al. 2018). Whether these findings are related to the differential cancer incidence between genders needs further investigation.

Tumour-specific SBGs were much less frequent, suggesting that once tumorigenesis starts the transcriptomic differences become more diluted between genders. Of notice, in GC we found the cancer-associated genes *LPCAT1* and *RAD51C* over-expressed in females (Bi et al. 2019; Somyajit, Subramanya, and Nagaraju 2010; Meindl et al. 2010). In TC, among over-expressed in males we found *PPARG,* previously reported in thyroid carcinomas (Raman and Koenig 2014), *ERCC5, MYH11, LEMD2, ZNF133 and IDH1,* the latter found to be mutated in thyroid carcinomas (H. Yang et al. 2012). *TFRC,* whose expression has been associated with poor prognosis and tumour progression (Shen et al. 2018), was found over-expressed in females. These genes are potential contributors to the gender-specific tumorigenesis of GC and TC.

Differential promoter methylation appears as a mechanism that might underlie some of the observed expression differences, given its incidence among the SBGs in both cancers and the tumour-normal differences observed for both genders in TC. Of the differentially methylated SBGs, most belong to the X-chromosome and are known to escape X-inactivation. Changes in promoter methylation in genes escaping X-inactivation have been previously found (Sharp et al. 2011). However, 46% and 80% of the SBGs in GC and TC were not differentially methylated in our study, belonging mostly to the autosomes. Moreover, half of the SBGs were not profiled by methylation probes. Other processes that contribute to sex-specific gene regulation include epigenetic regulation of enhancers (Reizel et al. 2015), estrogen-regulated miRNA expression (Klinge 2012) and tRNA regulatory fragments (Telonis et al. 2019). Whether these processes are involved in the sex-biased patterns not explained by differential promoter methylation remains an open question.

Network analysis of co-expressed genes may help in identifying gender-specific cellular rewirings in normal and tumour tissues. Our results show that most of the network modules are preserved between genders, in agreement with previous work (Melé et al. 2015), in both types of tissues. The reasons for the different degrees and modalities of gene module preservation, in tumour and normal tissues, remains to be studied in the future. Nevertheless, we found gene modules preferentially associated for males in GC and for females in TC, further associated with cancer histological subtypes.

The normal tissue surrounding the tumour can be influenced by pro-inflammatory signals released by tumours, representing an intermediate state between tumour-free healthy tissue and established neoplasms (Aran et al. 2017). Nonetheless, our analysis of the tumour-normal differential transcriptome found that a significant fraction of the DEGs were oncogenes, as previously found (Pranavathiyani et al. 2019). We also found that the most affected sex in each cancer shows an under-expression of TSGs in tumours, which is an important cancer hallmark (Hanahan and Weinberg 2011). The same result was not observed for the opposite sex, supporting the hypothesis that TSGs inactivation or deregulation may occur in the tumour tissues of the most susceptible gender. This result reinforces that genders may follow different carcinogenic programs, and therefore appropriated and differentiated therapeutic strategies may be considered (J. Ma, Malladi, and Beck 2016; Yuan et al. 2016; Buoncervello et al. 2017). In TC the female-specific tumour-normal DEGs over-expressed in normal tissues were involved in the same immune-related pathways as the male-specific DEGs over-expressed in tumours. This result may highlight some still unknown predisposing elements that make females more susceptible to TC.

In summary, we were able to identify the gender-specific expression landscape in normal and tumour tissues of thyroid and stomach. We expect that these results provide novel insights in the understanding of the gender-differential risk underlying these cancers.

2.5. Methods

Data collection

We obtained TCGA mRNA-seq and clinical data for gastric cancer (GC) and thyroid cancer (TC) tumour matched-normal samples. The data was downloaded from the Genomic Data Commons (GDC) data portal (portal.gdc.cancer.gov/) in reads per kilobase of exon model per million mapped reads (RPKM) and read counts formats, at the gene-level. Additional clinical information was obtained from (Bass et al. 2014; Agrawal et al. 2014; J. Liu et al. 2018). The TCGA methylation data was acquired from the FireBrowse portal (firebrowse.org/), as beta values per methylation probe (450k arrays). We also compiled GTEx v6 mRNA-seq and phenotypic data, for stomach and thyroid normal samples, from the GTEx portal (gtexportal.org/home/) in RPKM and read counts formats. The gene annotation was downloaded from the GDC portal for the TCGA mRNA-seq data (GENCODE v22) and from the GENCODE website (gencodegenes.org/) for the GTEx mRNA-seq data (GENCODE v19).

A list of human TSGs (Zhao, Sun, and Zhao 2013) and oncogenes (Yining Liu, Sun, and Zhao 2017) were downloaded from bioinfo.uth.edu/TSGene/ and ongene.bioinfo-minzhao.org/index.html, respectively. The Cancer Gene Census catalogue (Sondka et al. 2018) was downloaded from cancer.sanger.ac.uk/census. A list of cancer driver genes was downloaded from (M. H. Bailey et al. 2018). The X-chromosomal genes known to escape inactivation were obtained from (Tukiainen et al. 2017).

Data pre-processing

We assembled the TCGA mRNA-seq (read counts and RPKMs) and clinical data in tabular formats using in-house scripts. The datasets comprised 60483 genes across 407 samples (375 primary tumours and 32 matched-normal) for GC and 560 samples (502 primary tumours and 58 matched-normal) for TC. For downstream analysis we selected

only the protein-coding and long intervening/intergenic non-coding RNA (lincRNA) genes, comprising 27470 genes, as described in GENCODE v22 annotation. In order to remove lowly-expressed genes, we filtered out those without 5 counts-per-million (CPM) in at least 20% of the tumour or normal samples. After gene filtering, the mRNA-seq datasets comprised 12690 genes for TC and 13674 genes for GC.

The GTEx mRNA-seq datasets (read counts and RPKMs) comprised 56318 genes, across 193 samples for stomach and 323 samples for thyroid tissues. After selecting the protein-coding and lincRNA genes (27459 genes), as described in GENCODE v19 annotation, we removed those genes without 5 CPM in at least 20% of samples. The final mRNA-seq datasets comprised 12501 genes for thyroid and 12371 genes for stomach tissues. The CPM values were calculated using the *cpm* function from the edgeR package (Robinson, McCarthy, and Smyth 2009).

After merging the TCGA and GTEx samples in each tissue, the final mRNA-seq datasets comprised 11734 genes for thyroid and 11842 genes for stomach. A PCA analysis was then performed using *prcomp* function in R.

We regressed-out potential confounding covariates from the GTEx gene expression data (log2 RPKM) using the following multiple linear model:

Equation 2.1: $g_i = \beta_0 + \beta_1 \text{smrin} + \beta_2 \text{age} + \beta_3 \text{ethncty} + \beta_4 \text{mhcancernm} + \beta_5 \text{smcenter} + \beta_6 \text{smtstptref} + \beta_7 \text{smnabtcht} + \beta_8 \text{smtsisch} + \varepsilon$

where g_i represents the gene expression for gene i, β_0 the intercept, β_i i $\in$ (1, ..., 8), the regression coefficients for the covariates, and ε the noise term. See **table S2.8** for additional information about the covariates. The gene expression corrected for these covariates corresponded to the residuals of this model, calculated as:

Equation 2.2: $g_i' = g_i - \hat{g}_i$

where g_i' represents the gene expression corrected for these covariates, g_i the observed gene expression and $\hat{g}_i$ the predicted gene expression from the model. The linear models were calculated using the *lm* R function.

Differential gene expression

We performed the differential expression analyses using the edgeR package. edgeR models the variance of the read counts per gene using a negative binomial distribution and applies a generalized linear model (GLM) to account for additional covariates when testing for differential expression.

We performed differential gene expression between genders in TCGA tumour and GTEx normal samples from stomach and thyroid. We also performed differential gene expression between TCGA tumours and matched-normal samples in each gender. In each comparison we created a design matrix taking into account several covariates. In R notation:

Male vs female in TCGA tumour samples:
design = model.matrix(~ race + ethnicity + age + tumor stage + histology + tss + portion + plate + gender, data = covars.matrix)

Male vs female in GTEx normal samples:
design = model.matrix(~ smrin + age + ethncty + mhcancernm + smcenter + smtstptref + smnabtcht + smtsisch + gender, data = covars.matrix)

Tumour vs matched-normal TCGA samples in each sex:
design = model.matrix(~ race + ethnicity + age + tss + portion + plate + tissue type, data = covars.matrix)

where *design* corresponds to the design matrix*, model.matrix* the R function used to define the design matrices, each term (e.g. *age*) the respective covariate, and *covars.matrix* the data frame containing the covariates. See **Table S2.8** for additional information about the covariates. In each comparison we normalized the read counts using the trimmed-mean of M-values method (Robinson and Oshlack 2010), with the *calcNormFactors* function. After estimating the common and tagwise dispersions with *estimateDisp*, we fitted a GLM model for each gene using *glmFit*. A likelihood ratio test (LRT) was then applied on the coefficients of tissue type (tumour or normal) or gender (male or female) to test for differences between these samples, using the *glmLRT* function. The *P-values* were adjusted for false discovery rate using the Benjamini-Hochberg procedure.

We selected the differentially expressed genes between genders (sex-biased genes [SBGs]) using a *FDR* lower than 5%, additionally requiring for the differentially expressed genes (DEGs) between tumour and matched-normal samples an absolute log2 fold-change higher than 1. Tumour and normal-specific SBGs were calculated by intersecting both gene sets. The same process was performed to calculate male- and female-specific DEGs between tumour and matched-normal samples.

We also performed a differential expression analysis between genders using the TCGA normal samples, adjusted for race, ethnicity, age and portion (**Table S2.8**). In this analysis we found 11 and 26 SBGs in stomach and thyroid, respectively, and only two normal-specific SBGs in both tissues. The relative low number of SBGs, alongside the higher number of samples in the GTEx cohort, led us to consider the GTEx dataset for the further analyses in this paper. We rationale that the high number of samples in GTEx would help to robustly quantify the gender differential transcriptome.

We also used limma (with voom) differential expression method (Law et al. 2014), which assumes a normal distribution for the gene expression data, instead of a negative binomial distribution as edgeR. We found an overlap greater than 85% with edgeR (in all comparisons), indicating a robust set of called differentially expressed genes.

Differential gene promoter methylation

We selected TCGA methylation probes annotated to gene promoter regions using the R package IlluminaHumanMethylation450kanno.ilmn12.hg19. Then, for each gene in each sample, we calculated the average beta value across the promoter probes.

The differential gene promoter methylation analysis was performed using a Wilcoxon rank-sum test (*wilcox.test* R function). We assessed differences on gene promoter methylation between genders in TC and GC, and between tumour and matched-normal samples in TC, independently for each gender. The GC matched-normal samples were not profiled by the selected methylation array, hampering the tumour-normal differential methylation analysis in GC. The *P-values* were adjusted for false discovery rate using the Benjamini-Hochberg procedure. Genes with $FDR < 5\%$ were defined as differentially methylated.

In GC and TC, 56% of the genes with expression data were covered by methylation probes. For the SBGs, 51% in GC and 48% in TC contained information about the differential methylation status (**Supplementary figure 2.7A, Supplementary figure 2.7B**). For the tumour-normal DEGs in TC, 30% in females and 34% in males were profiled by differential methylation (**Supplementary figure 2.7C, Supplementary figure 2.7D**).

Construction of gene co-expression networks

We built gene co-expression networks for each gender in TCGA tumour and GTEx normal samples from stomach and thyroid, using the methods in the Weighted Correlation Network Analysis (WGCNA) package (Langfelder and Horvath 2008). First, we log2 transformed and transposed the RPKM matrices. For the GTEx samples we used the gene expression data after regressing-out the confounding covariates. Then, we removed potential outlier samples using a hierarchical clustering dendrogram (*hclust* R function, using *average* as agglomeration method). The distances between samples were calculated using the *euclidean* distance measure, using the *dist* function. In GC we removed 8 samples in females *(cut height = 117)* and 4 samples in males *(123)*. In TC we removed 3 samples in females *(120)* and 4 samples in males *(100)*, while in thyroid normal tissue we removed 3 samples in females *(70)* and 1 sample in males *(80)*.

Network construction:

In order to build gene co-expression networks, we calculated an adjacency matrix using the following expression:

Equation 2.3: $a_{ij} = |corr(g_i, g_j)|^{\beta}$

where a_{ij} corresponds to the connection strength between gene i (g_i) and gene j (g_j), |corr| the absolute Pearson's correlation coefficient, and β the soft-thresholding power that approximates the network of a scale-free topology. Raising the absolute correlation values to a power accentuates high correlations at the expense of low correlations.

Module detection:

After network construction the next step was to find gene modules or clusters of densely interconnected genes. For that, we converted the adjacency matrix into a topological overlap matrix (TOM). The TOM contains the pairwise relative interconnectedness of all nodes in the network. After converting the TOM into a dissimilarity measure (1 - TOM), we identified gene modules by cutting off the branches of a hierarchical clustering dendrogram (hclust R function using *average* as agglomeration method). The branches were cut using the Dynamic Hybrid algorithm (Langfelder, Zhang, and Horvath 2008). This method eliminates the need of using constant height cutoff values and is more effective in complex dendrograms. The gene expression profiles of a module can be summarized by its module eigengene (first principal component). We merged highly similar

modules if the eigengene Pearson's correlation was higher than 0.75. Genes without module assignment were not considered for further analyses.

WGCNA functions:

We performed the network construction and module detection steps automatically and sequentially, using the *blockwiseModules* function. We used the parameters as shown by the authors in the WGCNA tutorials (tutorial I, section 2.a.) (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/). The only exceptions were the parameters *maxBlockSize*, *power* and *minModuleSize*. In order to perform the network construction in a single gene block we increased the maximum block size (*maxBlockSize*) from 5000 (the default) to 20000. This prevents the network module assignments from being split into multiple blocks. We estimated the optimal powers that approximate the networks of a scale-free topology, using the *pickSoftThreshold* function. Following WGCNA recommendations, we selected the highest power that exceeds the scale-free topology fit R^2 cutoff, set to 0.85 by default. In thyroid the powers were set to 9 for all networks, except for the males network in tumours, whose power was set to 10. For stomach the powers were set to 4 for all networks. Instead of 30 genes as minimum module size (*minModuleSize*), we opted to keep the default value of 20 genes. Using this approach, we built 8 networks in total.

Gender differential co-expression network analysis

We compared male to female networks using a strategy based on (Melé et al. 2015). We started by computing the percentage of gene content overlap between each pair of modules, where each module belongs to a different network. As an example, given the modules M and F in the males and females network, respectively, we calculated the overlap between M and F as follows:

Equation 2.4: $overlap_{MF} = \frac{|M \cap F|}{min(M,F)} \times 100$

where $overlap_{MF}$ corresponds to the percentage of overlap between M and F, $|M \cap F|$ is the number of genes in common between M and F, and min(M, F) is the length of the smaller module. We also calculated a Fisher's exact test *P-value* for each overlap, using the

overlapTable function from the WGCNA package. Based on the *P-value* and on $overlap_{MF}$, we considered the overlap of a given pair of modules as:

- *absent*, if *P-value* > 5% or (*P-value* < 5% and $overlap_{MF}$ < 20%);
- *low*, if *P-value* < 5% and (20% ≤ $overlap_{MF}$ < 50%);
- *moderate*, if *P-value* < 5% and (50% ≤ $overlap_{MF}$ < 70%);
- *high*, if *P-value* < 5% and ($overlap_{MF}$ ≥ 70%).

For both genders the modules were classified as *lowly*, *moderately* or *highly*-preserved, if the overlap with the opposite gender has been defined as *low*, *moderate* or *high*, respectively. We honoured the highest overlap when multiple overlaps occurred. As an example, in the stomach normal networks from GTEx, the largest module from males (5976 genes) has a *high* and a *moderate* overlap with two modules from females (with 3903 and 936 genes, respectively) (**Supplementary figure 2.12B**). Therefore, we considered the male module as highly-preserved in females. Modules without overlap with the modules of the opposite gender were classified as gender-specific.

Association of gender-specific modules with cancer clinical traits

The biological significance of a module can be defined as the absolute correlation between the module eigengene *E* and a sample phenotype *P* (Langfelder and Horvath 2008). Modules with high biological significance (correlation) can represent pathways associated with the phenotype *P*. We evaluated the biological significance of the gender-specific modules in tumours by fitting a linear regression model as follows:

Equation 2.5: $E = \beta_0 + \beta_1 P + \varepsilon$

where E represents the module eigengene, β_0 the intercept, ε the noise term and P the cancer clinical traits overall survival (in days), tumour stage (American Joint Committee on Cancer [AJCC] staging system) and cancer histological subtype. In TC, we considered the cancer histological subtypes classical (number of samples: 257), follicular (75) and tall cell (26). In GC, the cancer subtypes signet ring (9), diffuse (36), intestinal mucinous (14), intestinal NOS (not otherwise specified) (45), intestinal papillary (3) and intestinal tubular

(39). The association between the modules and the clinical traits was then evaluated using the regression-derived R^2 and the one-way ANOVA *P-values*. The linear models were calculated using the *lm* R function. We also evaluated the association between survival and the modules eigengene using univariate cox proportional hazard regression models. These models were computed using the *coxph* function from the survival R package.

In a given module related with phenotypic traits, the hub genes (highly connected) are the most relevant genes to look for, since their expression profiles can represent that of the entire module (Langfelder and Horvath 2008). In the gender-specific modules associated with the cancer histological subtypes, we investigated the cancer subtypes where these genes are predominantly expressed. For that we selected the hub genes of each module (with absolute intramodular connectivity [$|K_{ME}|$] > 0.8) and tested them for differential expression between cancer histological subtypes, using a Kruskal-Wallis rank sum test (*kruskal.test* R function).

Functional enrichment analysis

We performed functional enrichment using hypergeometric tests and gene set enrichment analysis (GSEA), implemented in the functions *enrichr* and *GSEA* from the clusterProfiler R package (G. Yu et al. 2012). We used gene sets downloaded from the MSigDB database (software.broadinstitute.org/gsea/msigdb), including C1 positional sets, C2 KEGG pathways and C5 GO biological processes (BP). We applied GSEA on gene modules derived from gene co-expression networks, sorted by K_{ME}. The comparison of enrichment profiles between gender-specific DEGs over-expressed in tumour or normal tissues was performed using the hypergeometric test implemented in the function *compareCluster*. The *P-values* were adjusted for false discovery rate using the Benjamini-Hochberg procedure.

The enrichment for TSGs, oncogenes, X-chromosomal genes escaping inactivation and differentially methylated genes was performed using a Fisher's exact test *(fisher.test* R function; *alternative = "greater")*.

The backgrounds corresponded to all genes that were analysed in our study, either by differential expression or by gene co-expression networks.

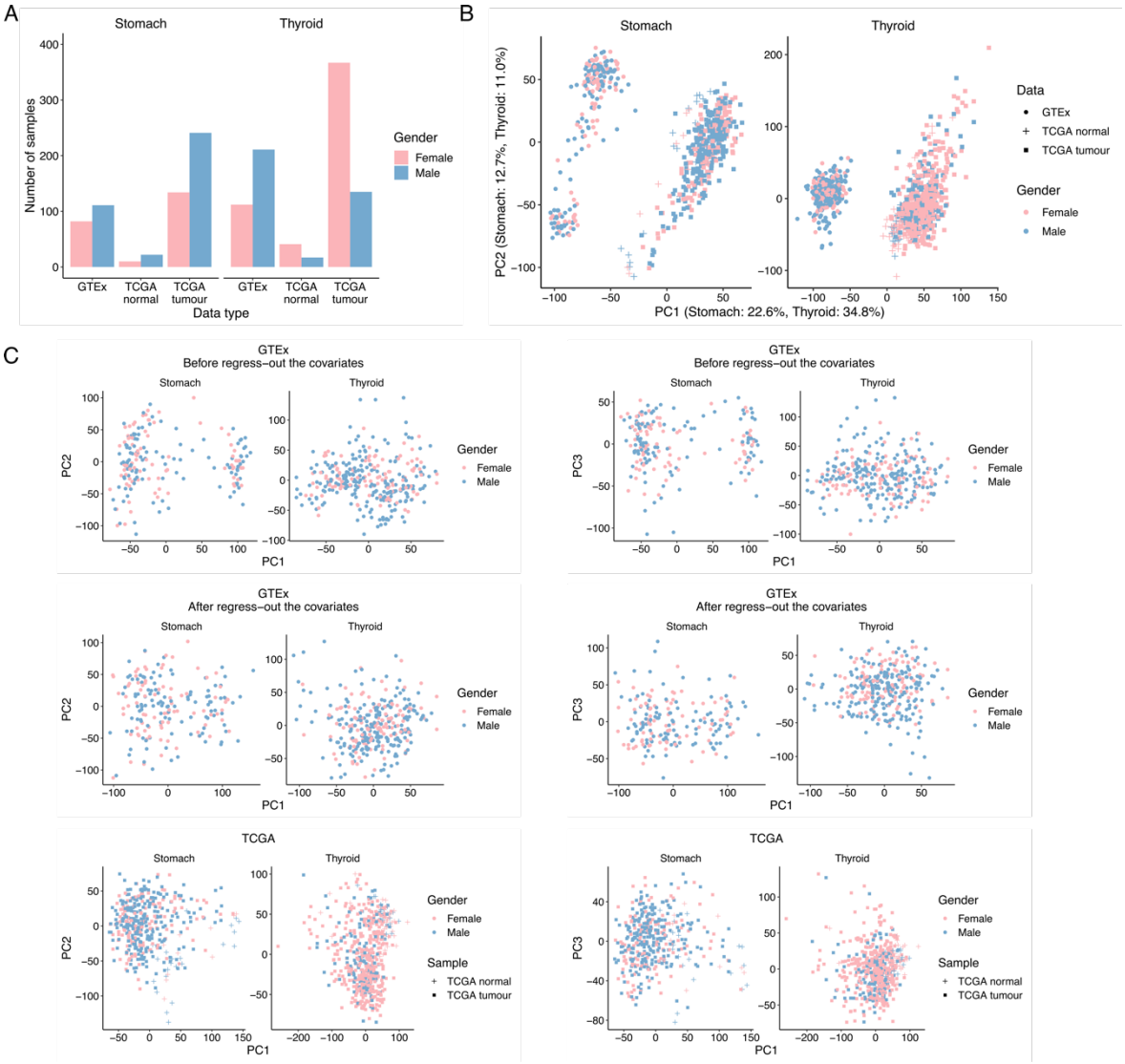
All gene lists reported in this study were annotated with functional gene summaries, using the function *queryMany* from the mygene R package.

Code availability

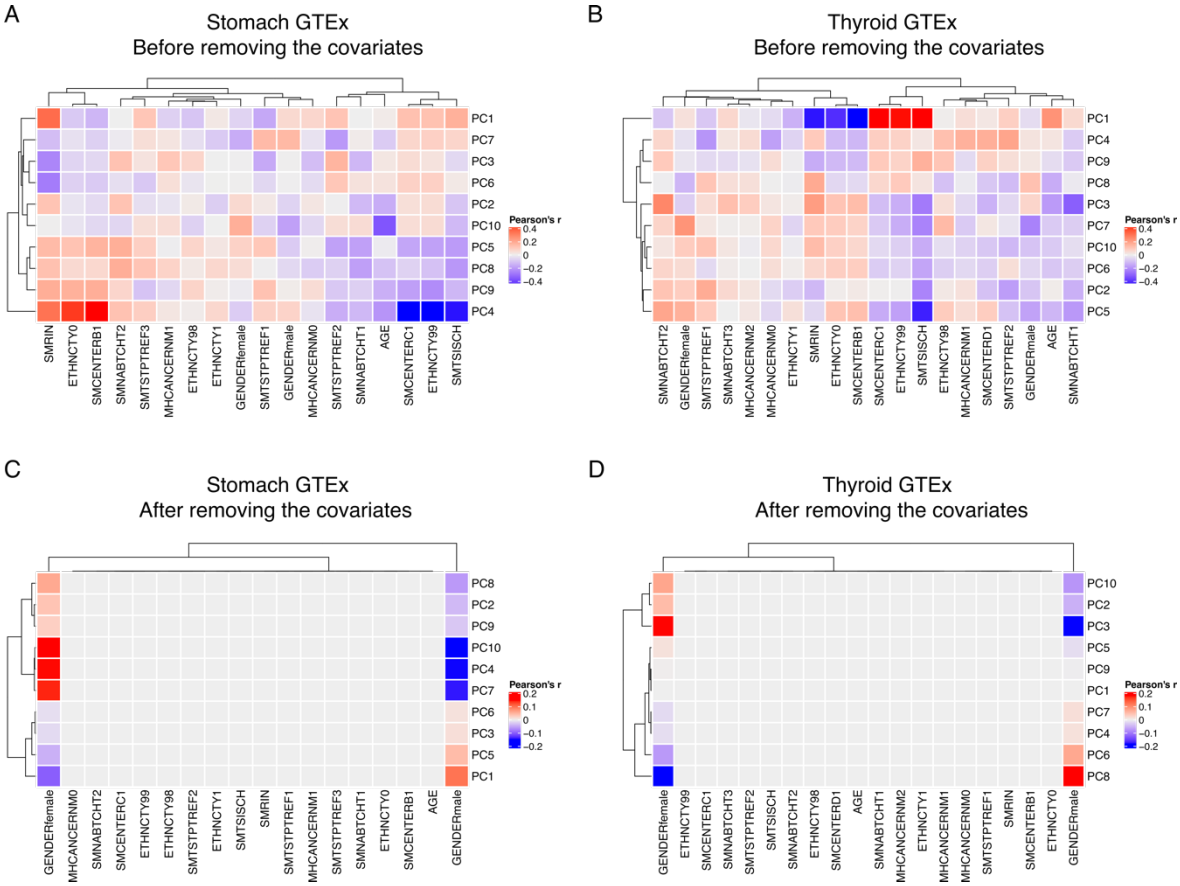
The computational analyses were performed in R 3.6.3 and all the code is available under a GNU General Public License V3 in a GitHub project, at the following url: github.com/abelfsousa/gender_differences. The differential expression analyses were performed with edgeR 3.26.8 and the gene co-expression network analyses with WGCNA 1.68. The functional enrichment analysis (hypergeometric tests and GSEA) were performed using clusterProfiler 3.12.0. Plotting was done using ggplot2 3.2.1, ComplexHeatmap 2.0.0, arcdiagram 0.1.12 and eulerr 6.1.0. Data analysis and structuring using dplyr 0.8.3, tidyr 1.0.0 and the remaining packages included in tidyverse 1.2.1.

2.6. Supplementary materials

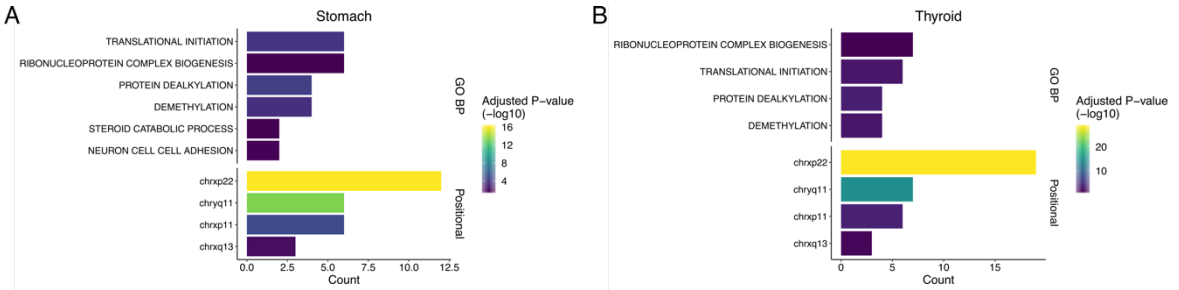
2.6.1. Figures



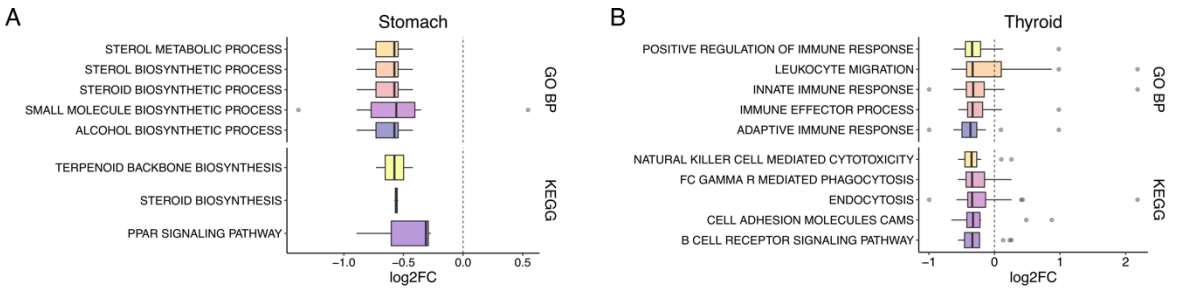
Supplementary figure 2.1. Number of samples and PCA analysis. (A) Number of samples by gender and data type (GTEx, TCGA tumour and TCGA normal) for stomach and thyroid. **(B)** Scatter plot representing the PCA analysis (each dot is a sample). The dot shape represents the data type and the color represents the gender. PC1 explains most of the variance and separates GTEx from TCGA samples. **(C)** Confounding effects regressed-out from the GTEx RNA-seq data (**Methods**). The top plots show the PC1 vs PC2 (left) and PC1 vs PC3 (right) before regressing-out the covariates. The middle plots show the same data after regressing-out the covariates. The TCGA RNA-seq data (bottom plots) was not regressed-out, as the PCA did not reveal a clear sample separation.



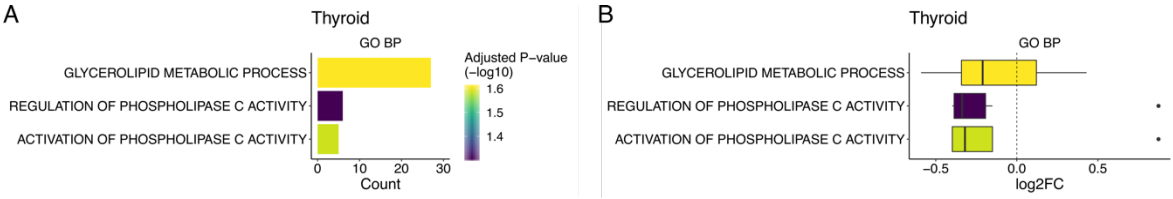
Supplementary figure 2.2. Pearson correlation of the first 10 principal components (PCs) from PCA with covariates before and after normalization. Stomach and thyroid before **(A) (B)** and after removing the covariate effect **(C) (D)**. The PCA was performed using the GTEx gene expression data (log2 FPKM). Categorical covariates were converted into binary variables. Each level is represented in the heatmap (columns). The gender covariate was not regressed-out. See **Table S2.8** for information about the covariates and **Methods** for details about the model used.



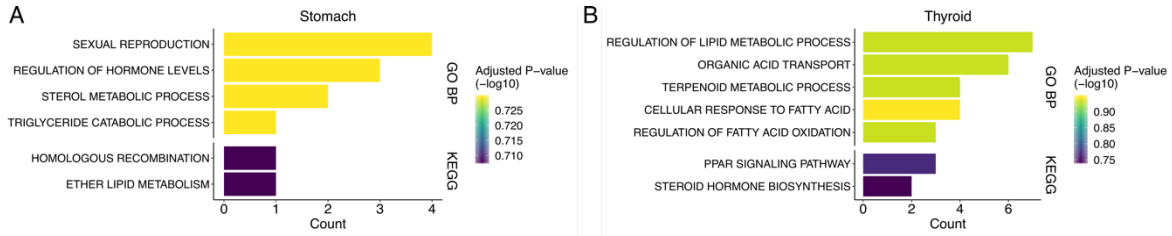
Supplementary figure 2.3. Functional enrichment analysis on the SBGs shared by the normal and tumour tissues. GO biological processes (GO BP) and genomic positions (Positional) enriched in stomach **(A)** and in thyroid **(B)** (top 5; FDR < 5%).



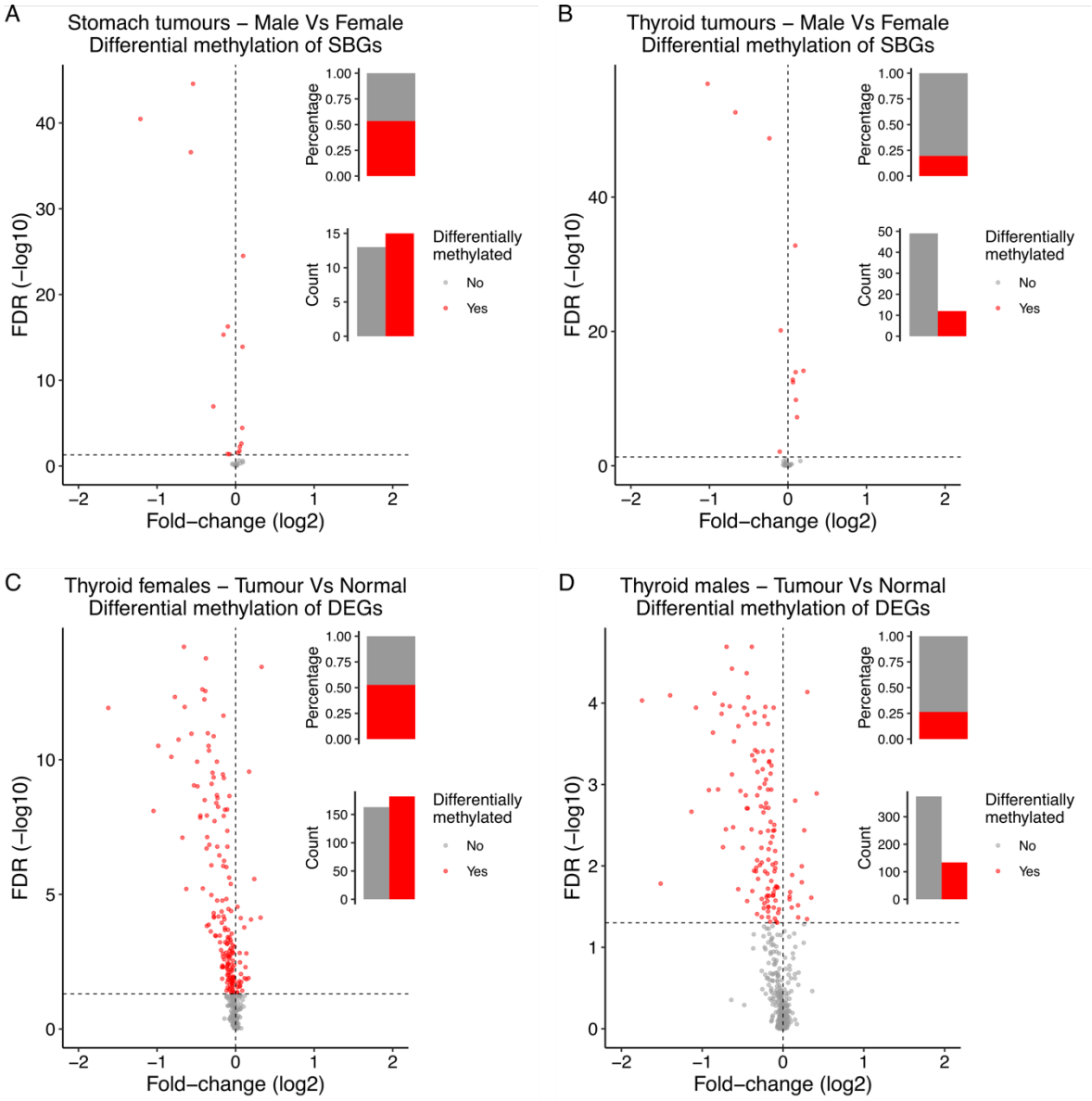
Supplementary figure 2.4. Distribution of gene fold-changes (log2) for the enriched terms in the normal-specific SBGs (related to the main Figure 2.2G and Figure 2.2H). **(A)** is for stomach and **(B)** is for thyroid. To the left of the vertical lines are the genes over-expressed in females and to the right the genes under-expressed in females (over-expressed in males).



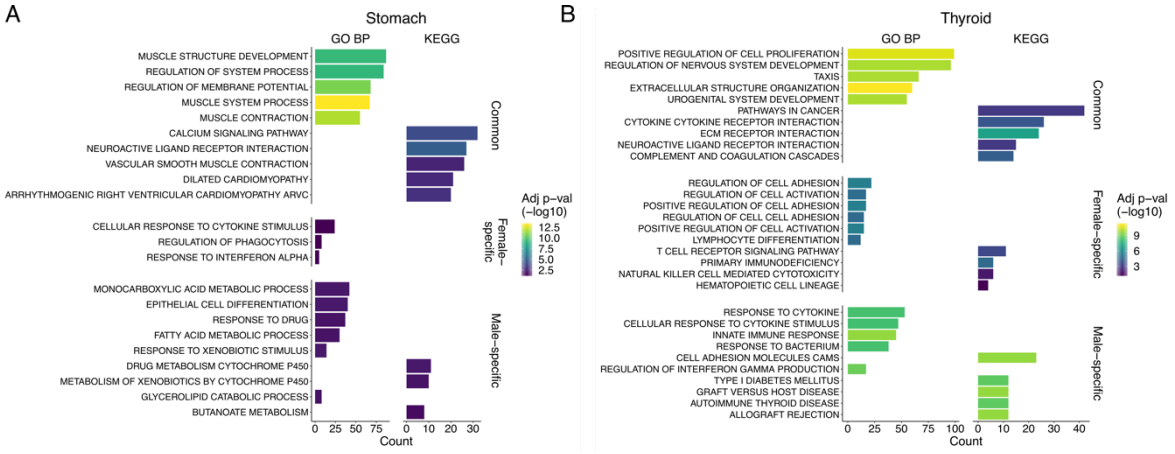
Supplementary figure 2.5. GO biological processes (GO BP) related to lipids metabolism enriched in the thyroid normal-specific SBGs. (A) GO BP terms enriched (FDR < 5%). **(B)** Distribution of gene fold-changes (log2) for the enriched terms. To the left of the vertical lines are the genes over-expressed in females and to the right the genes under-expressed in females (over-expressed in males).



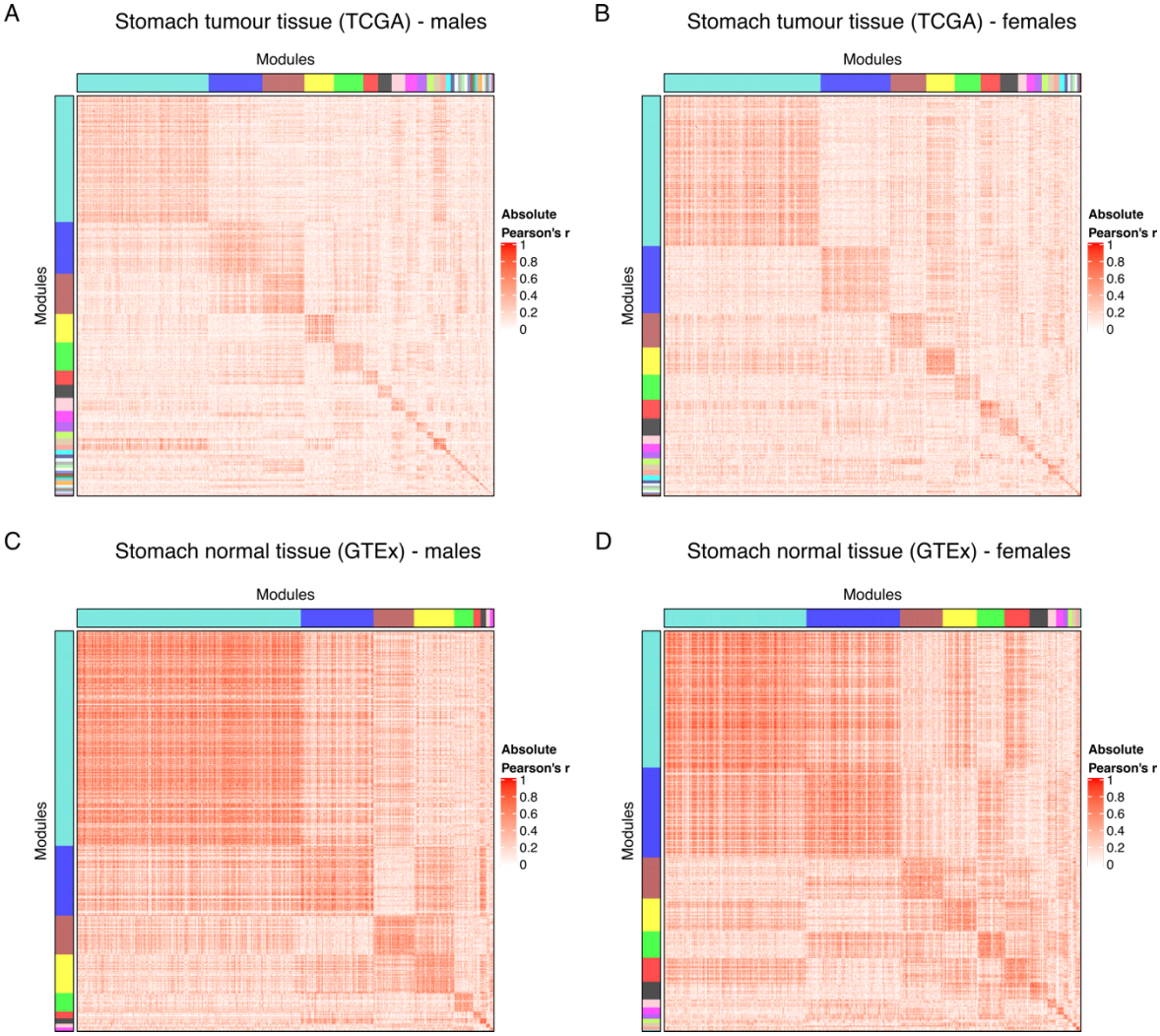
Supplementary figure 2.6. Functional enrichment analysis on the tumour-specific SBGs. GO biological processes (GO BP) and KEGG pathways enriched in stomach **(A)** and in thyroid **(B)** (top 5; FDR < 20%).



Supplementary figure 2.7. Differential promoter methylation analysis. Only the differentially expressed genes (gender-biased [SBGs] and tumour-normal [DEGs]) with information on differential methylation status are represented. (A) (B) Differential methylation status of the SBGs in GC and TC, respectively. Genes with FDR < 5% were defined as differentially methylated. **(C) (D)** Differential methylation status of the tumour-normal DEGs in TC for females and males, respectively. Genes with FDR < 5% were defined as differentially methylated. No methylation data (beta values per probe) was available for the tumour-matched normal samples of GC to perform the same analysis. The barplots in the middle and upper right corners show the number and percentage of differentially methylated genes, respectively. All differentially expressed gene sets showed an enrichment for differentially methylated genes in comparison to the background (genes not differentially expressed and with information on differential methylation status). P-value = 2.0e-15, 2.7e-6, 1.0e-3 and 1.7e-24, for (A), (B), (C) and (D), respectively.

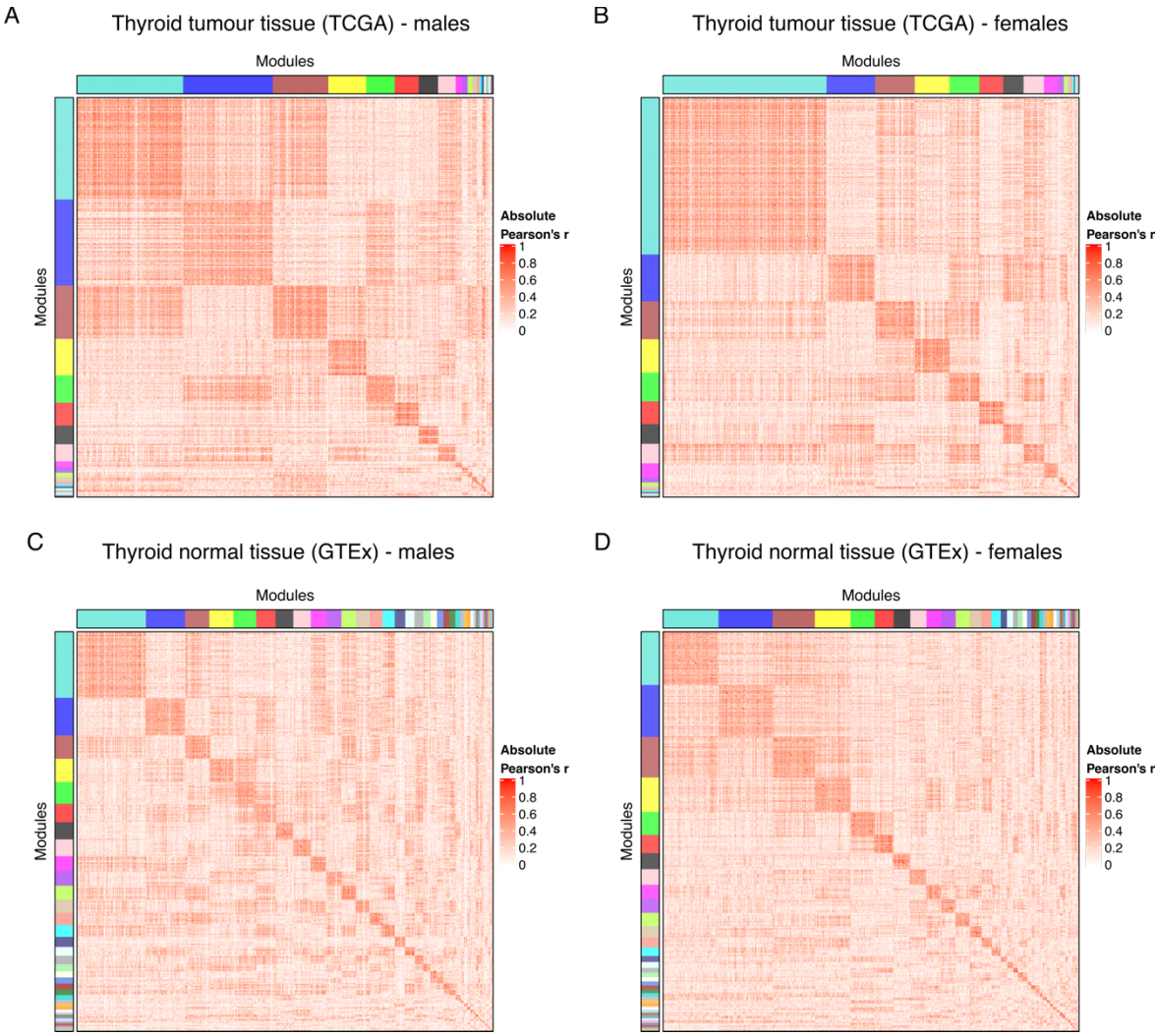


Supplementary figure 2.8. Functional enrichment analysis on the tumour-normal DEGs. GO biological processes (GO BP) and KEGG pathways enriched in stomach **(A)** and in thyroid **(B).** The terms are represented by DEG group: common, female-specific and male-specific (top 5; FDR < 5%).

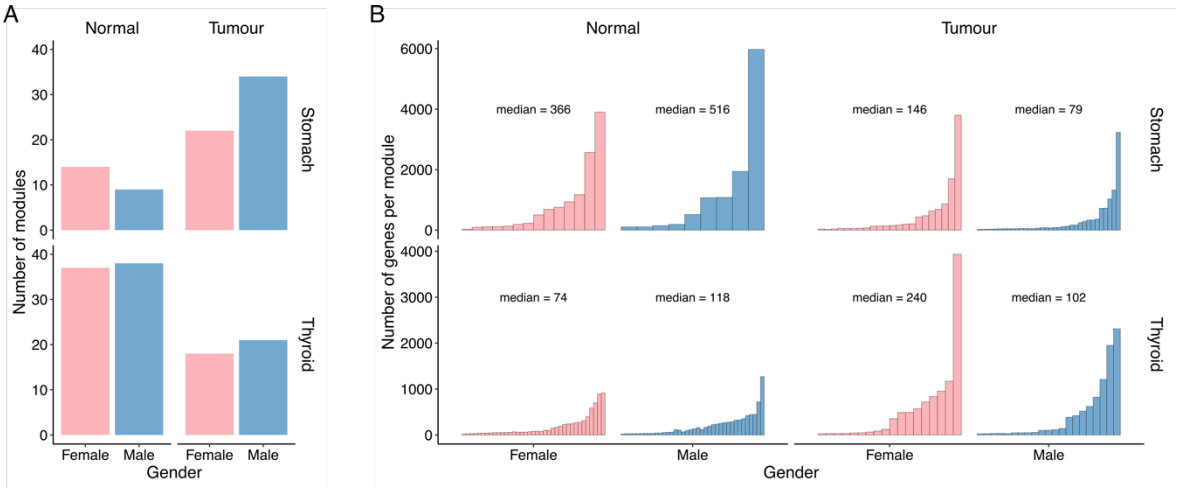


Supplementary figure 2.9. Correlation matrix of the WGCNA networks for the stomach tissue. Absolute Pearson correlation between gene expression profiles (log2 FPKM). The gene module assignments are shown by the colour bars at the top and left of

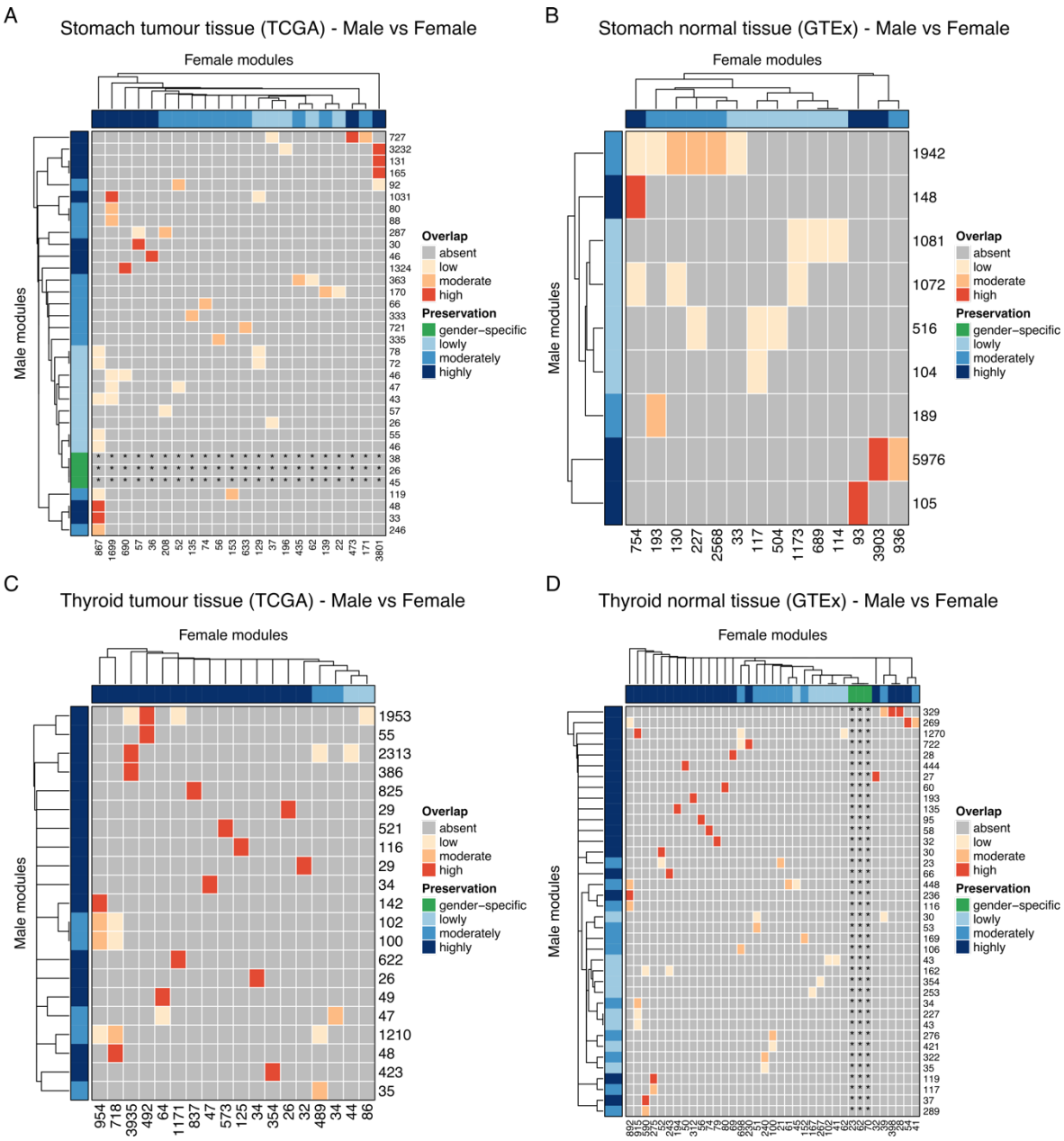
each plot. Network representation for the tumour tissue of males **(A)** and females **(B)** and for the normal tissue of males **(C)** and females **(D)**.



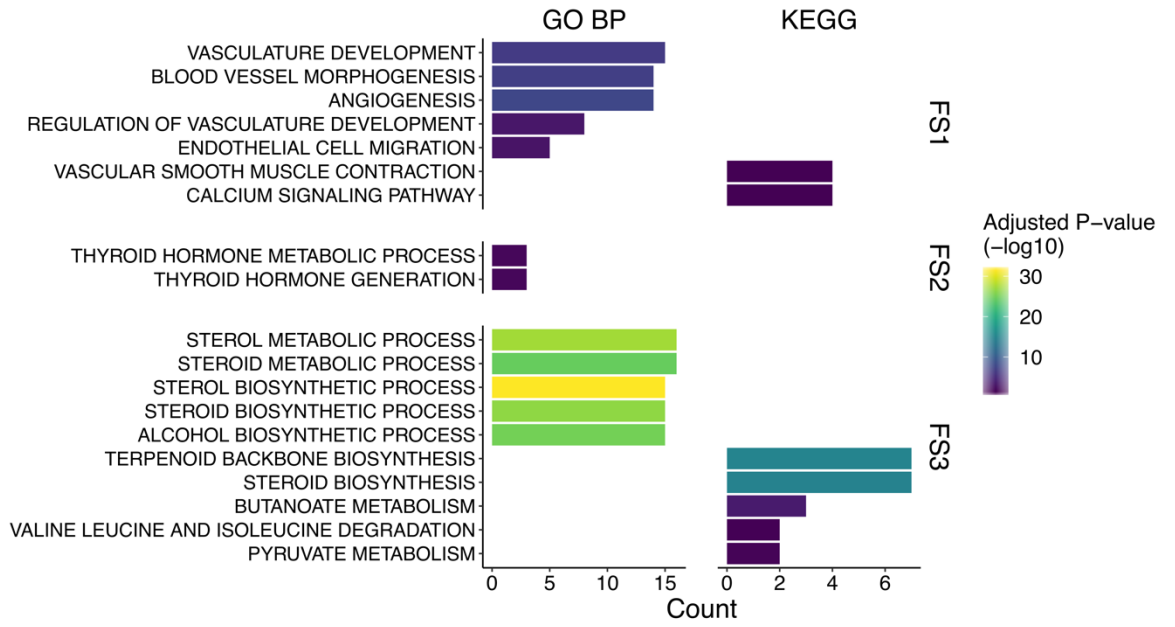
Supplementary figure 2.10. Correlation matrix of the WGCNA networks for the thyroid tissue. Absolute Pearson correlation between gene expression profiles (log2 FPKM). The gene module assignments are shown by the colour bars at the top and left of each plot. Network representation for the tumour tissue of males **(A)** and females **(B)** and for the normal tissue of males **(C)** and females **(D)**.



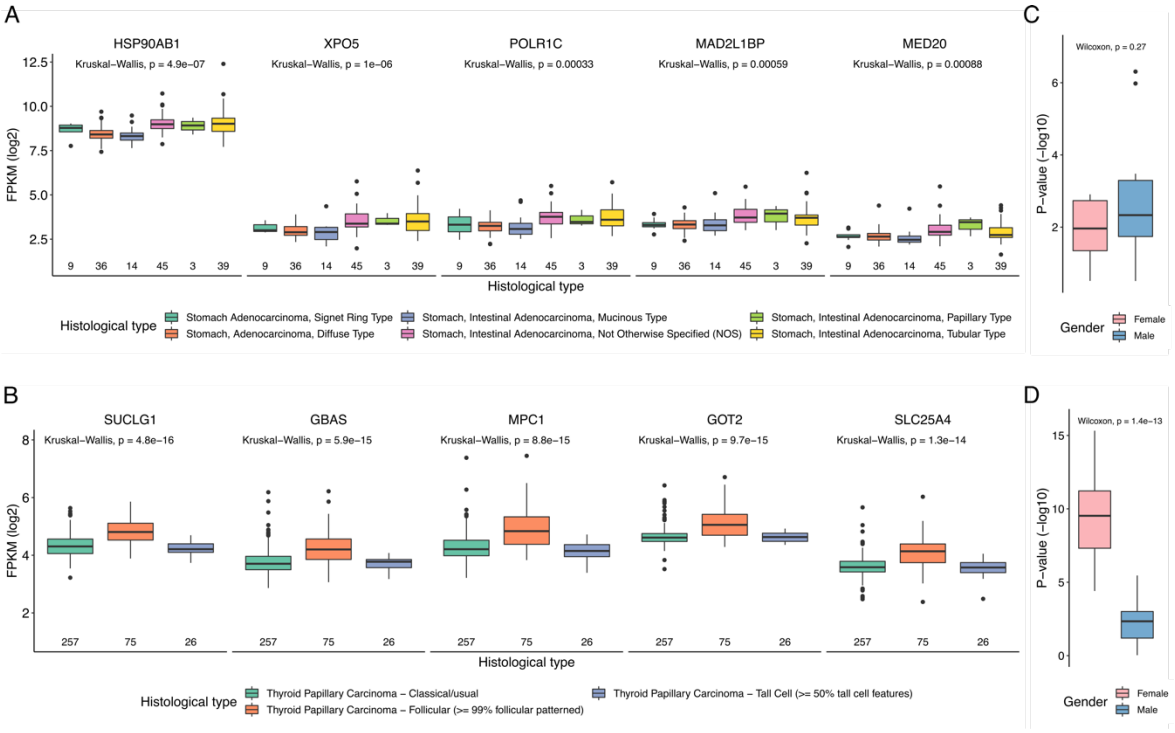
Supplementary figure 2.11. Features of gender co-expression networks in normal and tumour tissues of stomach and thyroid. (A) Number of gene modules in each network. **(B)** Number of genes across gene modules.



Supplementary figure 2.12. Comparison of modules between genders in the stomach and thyroid tissues. (A) and **(B)** are for the tumour and normal tissues of stomach; **(C)** and **(D)** are for the tumour and normal tissues of thyroid. Male modules are in the rows and female modules are in the columns. The numbers in the rows and columns are the number of genes inside each module. The Overlap legend represents the degree of gene content overlap between genders (**Methods**). The Preservation legend indicates the preservation of the modules in the opposite gender, based on the degree of overlap. Gender-specific modules are marked with asterisks.



Supplementary figure 2.13. Functional enrichment analysis of the female-specific modules in thyroid normal tissues. GO biological processes (GO BP) and KEGG pathways. The terms are represented for the female-specific modules FS1, FS2 and FS3 (top 5; FDR < 5%).



Supplementary figure 2.14. Differential expression between cancer histological subtypes of hub genes from gender-specific modules. (A) (B) Hub genes differentially expressed between cancer histological subtypes, in males for GC and in females for TC, respectively (top 5; P-value < 0.05). The number of samples is underneath each box. The Kruskal-Wallis rank sum test P-value is shown. **(C) (D)** Comparison of the Kruskal-Wallis P-value distribution between genders. The Wilcoxon rank-sum test P-value is shown. The

median of the distribution is higher for males in GC and for females in TC, the genders where the modules tend to be specific.

2.6.2. Tables

Supplementary table 2.1. Differentially expressed genes between males and females in tumour (TCGA) and normal (GTEx) stomach tissues.

Supplementary table 2.2. Differentially expressed genes between males and females in tumour (TCGA) and normal (GTEx) thyroid tissues.

Supplementary table 2.3. Number of X-chromosome genes escaping inactivation by SBG type in stomach and thyroid.

Supplementary table 2.4. Differentially expressed genes between stomach tumour and matched-normal tissues (TCGA) in each gender.

Supplementary table 2.5. Differentially expressed genes between thyroid tumour and matched-normal tissues (TCGA) in each gender.

Supplementary table 2.6. Differential expression between histological subtypes (in males) of hub genes from the male-specific module in GC.

Supplementary table 2.7. Differential expression between histological subtypes (in females) of hub genes from the female-specific module in TC.

Supplementary table 2.8. Description of covariates used in differential expression analyses.

All supplementary tables can be consulted using the following DOI:
doi.org/10.3389/fgene.2020.00808

3. Multi-Omics Characterization of Interaction-Mediated Control of Human Protein Abundance Levels

This chapter includes published material from the following article:
Abel Sousa, Emanuel Gonçalves, Bogdan Mirauta, David Ochoa, Oliver Stegle and Pedro Beltrão. Multi-omics Characterization of Interaction-mediated Control of Human Protein Abundance levels. Molecular & Cellular Proteomics, Volume 18, Issue 8, Pages 114-125, 9 August 2019.

3.1. Abstract

Proteogenomic studies of cancer samples have shown that copy-number variation can be attenuated at the protein level, for a large fraction of the proteome, likely due to the degradation of unassembled protein complex subunits. Such interaction-mediated control of protein abundance remains poorly characterized. To study this, we compiled genomic, (phospho)proteomic and structural data for hundreds of cancer samples and found that up to 42% of 8,124 analyzed proteins show signs of post-transcriptional control. We find evidence of interaction-dependent control of protein abundance, correlated with interface size, for 516 protein pairs, with some interactions further controlled by phosphorylation. Finally, these findings in cancer were reflected in variation in protein levels in normal tissues. Importantly, expression differences due to natural genetic variation were increasingly buffered from phenotypic differences for highly attenuated proteins. Altogether, this study further highlights the importance of post-transcriptional control of protein abundance in cancer and healthy cells.

3.2. Introduction

Cancer cells can harbour a large number of somatic DNA alterations ranging from point mutations to gene copy changes that can occur from deletion or amplification of small regions or whole chromosomes. While these events are the source of the genetic variation that can confer a selective advantage and lead to cancer, large changes in gene numbers can be detrimental and cause imbalances in the corresponding protein levels. Several studies have shown that the majority of changes in gene copy-number will propagate to changes in the corresponding protein levels (Dephoure et al. 2014; Stingele et al. 2012; Pavelka et al. 2010). However, models of aneuploidy of different species and analysis of gene copy-number variation (CNV) in cancer have shown that CNVs of protein coding genes belonging to protein complexes tend to be attenuated at protein level (Dephoure et al. 2014; Gonçalves et al. 2017; Ishikawa et al. 2017). In addition, we have shown that some complex members can act as rate-limiting subunits and indirectly control the degradation level of attenuated complex members (Gonçalves et al. 2017). These results are in-line with pulse chase degradation measurements showing that several complex subunits have a two-state degradation profile, that is compatible with a model in which they are expressed above the required levels and have a higher degradation rate when unbound from the complex

(McShane et al. 2016). The attenuation of changes at the protein level also justifies why protein complex subunits show higher correlation of protein abundances than the corresponding mRNA levels (Ryan et al. 2017; J. Wang et al. 2017), and why correlation analysis can be used to identify cancer-specific interaction networks (Lapek et al. 2017; Roumeliotis et al. 2017).

These results support a long-standing view that protein complex formation can set the total amount of protein levels (Abovich et al. 1985). The degradation of unbound subunits may be due to a requirement of avoiding free hydrophobic interface surfaces that can be prone to aggregate (Young, Jernigan, and Covell 1994). In eukaryotic species, this appears to be achieved by degrading excess production, while in bacterial and archaeal species genes coding for protein complex subunits tend to occur within operon structures such that they will be expressed at similar levels (Mushegian and Koonin 1996). This link between appropriate expression and complex formation is further emphasized by the preferential ordering of subunits in operons starting from the subunits that tend to assemble first (Wells, Bergendahl, and Marsh 2016).

While this phenomenon of gene dosage attenuation in protein complexes has been well documented, we still do not understand (i) what protein properties are associated with the propensity for a protein to be attenuated, (ii) nor if the characteristics of the attenuation process are seen in non-cancerous cells. Here we have extended on a previous analysis (Gonçalves et al. 2017), performing a multi-omics study of protein level attenuation of gene dosage that combines genomics, (phospho)proteomics and structural data. Analysing 8,124 genes/proteins we observed that up to 42% of proteins showed evidence of post-transcriptional regulation. Over 500 protein-protein interactions showed indirect control of degradation of one subunit via physical associations, 32 of which may be further controlled by phosphorylation. Using structural models for 3,082 interfaces, we found that a higher fraction of interface residues is associated with a higher degree of attenuation. Finally, we studied the impact of these findings on non-cancerous systems. We found that the protein interaction-mediated control of protein abundance has an impact on the variation of protein levels across tissues, and that the degree of attenuation correlates with the probability that natural variation with an impact on gene expression may result in a phenotypic consequence.

3.3. Results

3.3.1. Protein level attenuation of gene dosage associates with distinct essentiality and structural features

In order to study protein post-transcriptional control, we collected matched gene copy-number, mRNA and protein expression cancer datasets made available by The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumour Analysis consortium (CPTAC), for breast (BRCA) (Mertins et al. 2016; Koboldt, Fulton, et al. 2012), ovarian (HGSC) (H. Zhang et al. 2016; Bell et al. 2011) and colorectal (COREAD) cancers (B. Zhang et al. 2014; Muzny et al. 2012). In addition, we compiled existing protein/gene expression and copy-number data for cancer cell lines from Lapek et al. (BRCA) (Lapek et al. 2017), Roumeliotis et al. (COREAD) (Roumeliotis et al. 2017) and Lawrence et al. (BRCA) (Lawrence et al. 2015). In total, 368 cancer samples (294 tumours and 74 cell lines) were compiled in our study with matched gene expression, copy-number and protein abundance (**Figure 3.1A**). Principal component analysis (PCA) revealed the presence of confounding effects in the RNA and protein expression data (**Supplementary figure 3.1A, Supplementary figure 3.2A**). These effects are related to cancer type, experimental batch, type of proteomics experiment, and also patient gender and age. Therefore, these potential confounding effects were regressed-out from the RNA and protein expression data (**Methods**). After correction, the association between the principal components and the potential confounding effects was removed (**Supplementary figure 3.1B, Supplementary figure 3.2B**). In the combined dataset, the average mRNA-protein correlation is 0.44, which is in agreement with previous studies.

We then investigated the impact of CNV in cancer proteomes, using the strategy reported in Gonçalves et al. (Gonçalves et al. 2017). Due to the sparseness of the protein data, we selected genes with protein measurements in at least 25% of the 368 samples, comprising 8,124 genes with CNV, mRNA and protein expression. We included the CNV, mRNA and protein measurements for the 8,124 genes in the **Table S3.1**. For each gene, we then calculated the Pearson correlation coefficient between the CNV and the mRNA and the CNV and the protein, across samples. In order to assess the disagreement between the transcriptome and proteome regarding the copy-number changes, we calculated an attenuation potential, corresponding to the difference between Pearson coefficients (**Methods**). A higher attenuation potential suggests genes that have CNVs buffered at the protein level. As previously, we then clustered the genes by attenuation potential using an unsupervised Gaussian mixture model (GMM). Using this strategy, we identified 3,435 (42%) genes as attenuated at the protein level (2,578 low-attenuated and 857 high-attenuated) and 4,689 as non-attenuated (**Figure 3.1B**, **Figure 3.1C; Table S3.2**). These results indicate that up to 42% of genes show signs of gene dosage buffering at the protein

level, probably due to a post-transcriptional control of protein degradation, and robustly recapitulates previous findings on a smaller set of 6,418 genes (Gonçalves et al. 2017). To discard the possibility that the lack of correlation between CNV and protein could be due to noisiness in measuring protein levels, we asked if the correlation of protein or mRNA measurements, across samples, could predict known protein-protein associations from CORUM, among protein pairs at different levels of attenuation (**Supplementary figure 3.3**). As can be seen by the receiver operating characteristic (ROC) curves, protein interaction pairs are better predicted by protein than mRNA correlations in all classes, with the difference increasing with the attenuation level. If the attenuation was mostly explained by noise in the protein level measurements, then the opposite trend would be expected, where correlation of protein measurements would be noisier for highly attenuated proteins and a worse predictor of protein-protein interactions. These results indicate that attenuated genes are not defined as such because they have noisier protein measurements.

In line with previous findings, the list of attenuated genes is strongly enriched in well characterized protein complex members, and notably in members of large complexes (**Supplementary figure 3.4**). More, the attenuation potential is correlated with the number of subunits in a protein complex, indicating that members of large complexes have higher attenuation than those of small complexes (**Supplementary figure 3.4E**). Attenuated genes are also expected to show increased ubiquitination after proteasome inhibition, which was confirmed here using previously published data with 3 different proteasome inhibitors - MG-132, epoxomicin and bortezomib (Gonçalves et al. 2017; S. A. Wagner et al. 2011; Udeshi et al. 2013) (**Supplementary figure 3.5A**). Having defined a comprehensive list of genes/proteins with different degrees of attenuation, we then set out to characterize their physical and genetic properties.

We first asked if the level of attenuation relates to distinct essentiality features, based on gene essentiality defined by CRISPR-Cas9 screens (29083409). Highly-attenuated proteins showed higher gene essentiality than low- and non-attenuated proteins (**Supplementary figure 3.5B**) (Wilcoxon rank-sum test *P-value < 2.2e-16*, highly- vs non-attenuated proteins). This result is likely to be driven by the enrichment of protein complex members of essential complexes, such as the ribosome and spliceosome. We then studied the physical characteristics of these proteins such as length and structural properties. We found that the highly-attenuated proteins tend to have a smaller size (**Figure 3.1D**) (Wilcoxon rank-sum test *P-value < 2.2e-16*; highly- vs non-attenuated proteins), suggesting a size-dependent buffering mechanism. For the structural analysis, we considered a total of 2,392 proteins having structurally defined interface models (Mosca, Céol, and Aloy 2013). We illustrate this analysis with the COP9 signalosome complex (**Figure 3.1E**), where we noticed a trend in which the subunits with a larger surface buried in interfaces had the

strongest attenuation (**Figure 3.1F**). While the trend on a single complex is not significant (**Figure 3.1F**), this trend was supported across all proteins, with the average fraction of residues at interfaces increasing from the non-attenuated to the highly-attenuated proteins, in a statistically significant manner (**Figure 3.1G**).

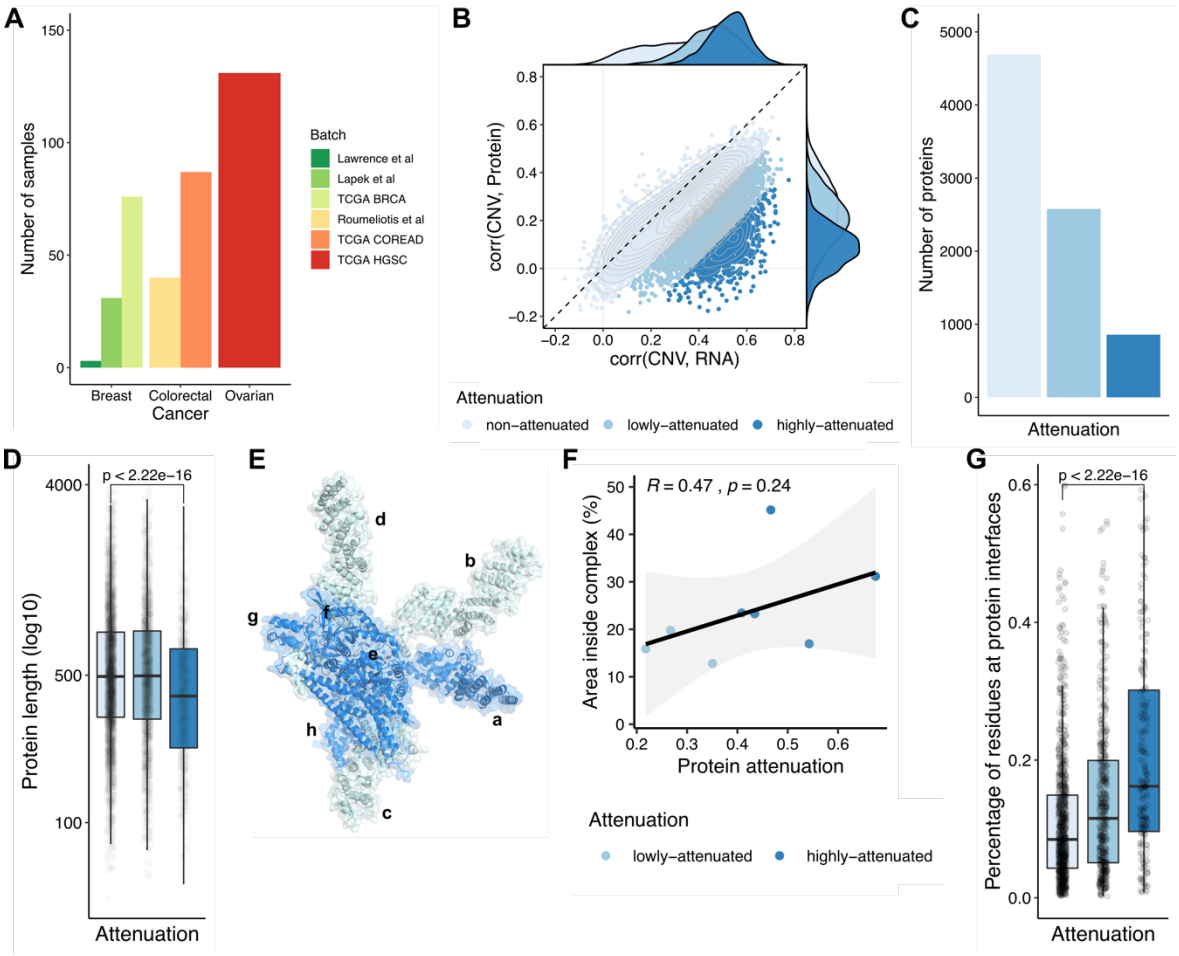


Figure 3.1. Features of proteins showing gene dosage buffering at the protein level. **(A)** Number of samples with CNV, mRNA and protein measurements, by cancer type and batch. **(B)** Scatter plot representing the correlation between the CNV and mRNA (x-axis) and the CNV and protein (y-axis), for each gene. The colours represent the attenuation levels. From light blue to dark blue: non-attenuated, lowly-attenuated and highly-attenuated. **(C)** Number of proteins by attenuation level. **(D)** Protein length (log10 of number of residues) by attenuation level. **(E)** Representation of COP9 signalosome complex. **(F)** Scatter plot representing the correlation between the attenuation potential (x-axis) and the fraction of residues at interfaces in the complex (y-axis), for the 8 protein subunits from the COP9 signalosome complex represented in **(E)**. **(G)** Percentage of residues at protein interfaces by attenuation level.

3.3.2. Protein interaction-dependent control of degradation depends on interface size

The features of highly attenuated proteins suggest that protein interactions are an important determinant of a protein's susceptibility of having gene dosage attenuation. It has been suggested that some members of protein complexes can act as scaffolding or rate-limiting subunits. We have previously analysed a set of 58,627 protein interactions among complexes curated in the CORUM database and identified a set of 48 interactions in which a protein can indirectly control the abundance of an interacting partner (Gonçalves et al. 2017). Here we set out to expand this analysis to all currently reported human physical interactions in the BioGRID database (**Methods**). In total, we collected 572,856 physical interactions and identified proteins whose CNV changes correlate with the protein abundance of interacting proteins once their mRNA levels are taken into account (**Methods**). For an interaction pair of proteins X and Y, we used a linear regression model, where we predict the protein levels of protein Y using the CNV of X, discounting the mRNA of Y and the impact of other covariates (**Methods**). Correlating molecular changes with DNA variation, such as CNVs, ensures the correlations found are most likely causal and in the direction of DNA changes to the molecular changes. Copy-number alterations in cancer most often occur in large segments leading to co-amplification or co-deletion of multiple co-localized genes. For proteins with two or more interacting partners that are genomically co-localized, we selected only the top-ranking association to avoid spurious "passenger" associations (**Methods**).

Out of 572,856 physical interactions, we had data to test associations for 411,591 with this model, finding 516 protein-protein associations as significant using CNV and mRNA (*false discovery rate [FDR]* < *5%*) (**Figure 3.2A; Table S3.3**). In this set of associations, we classified the proteins as *controlling* (353) - those capable of controlling the protein levels of their interacting partners; *controlled* (423) - whose abundance levels depend on their interactions; and *both* (60), as the proteins with the two characteristics (**Figure 3.2B**). Out of 423 *controlling* proteins, 62 had at least two interactions. The top *controlling* protein was TCP1, which was predicted to control the protein abundance of 7 complex partners, including CCT3, CCT5, CCT7 and CCT8 (**Figure 3.2D**). As expected, the *controlled* proteins had higher attenuation potential, a consequence of the post-transcriptional regulation of their protein levels (**Figure 3.2C**) (Wilcoxon rank-sum test *P-value* < *4.8e-6*; *controlled* vs *controlling* proteins). The *controlled* proteins also show a smaller size (Wilcoxon rank-sum test *P-value* < *9.8e-6*; *controlled* vs *controlling* proteins), which corroborates the hypothesis that protein size is important for the buffering mechanism

(**Figure 3.2E**). These results increased the evidence of interactions and regulators that may act as drivers of protein complex assembly.

We hypothesized that protein interaction-dependent control of degradation could depend on the protein interfaces size. To test this, we identified 60 significant associations with available structural models (**Methods**) and correlated the protein interface size with the effect-size (beta value) and significance (FDR) of the respective protein association pairs (**Figure 3.2F**). We found that both statistics are positively and significantly correlated with interface size (CNV beta - *Pearson's r = 0.36*, *P-value = 4.5e-3*; -log10 *FDR* - *Pearson's r = 0.46*, *P-value = 2.0e-4*). We selected two examples to illustrate the observed differences (**Figure 3.2G**). Post-transcriptional regulation of TRMT61A by TRMT6, that form the tRNA (adenine-N1-)-methyltransferase enzyme, is the second strongest association found in our analysis, and the interface formed between these two proteins covers a total of 72 residues. In contrast, a weaker association between CSNK2A1 and CSNK2B may be explainable by a much smaller interface of 13 residues.

These results show that interface sizes are an important determinant of the protein interaction mediated control of protein degradation. This may be due to an effect of binding affinity or differences in the recognition of exposed interfaces of different sizes by the degradation machinery.

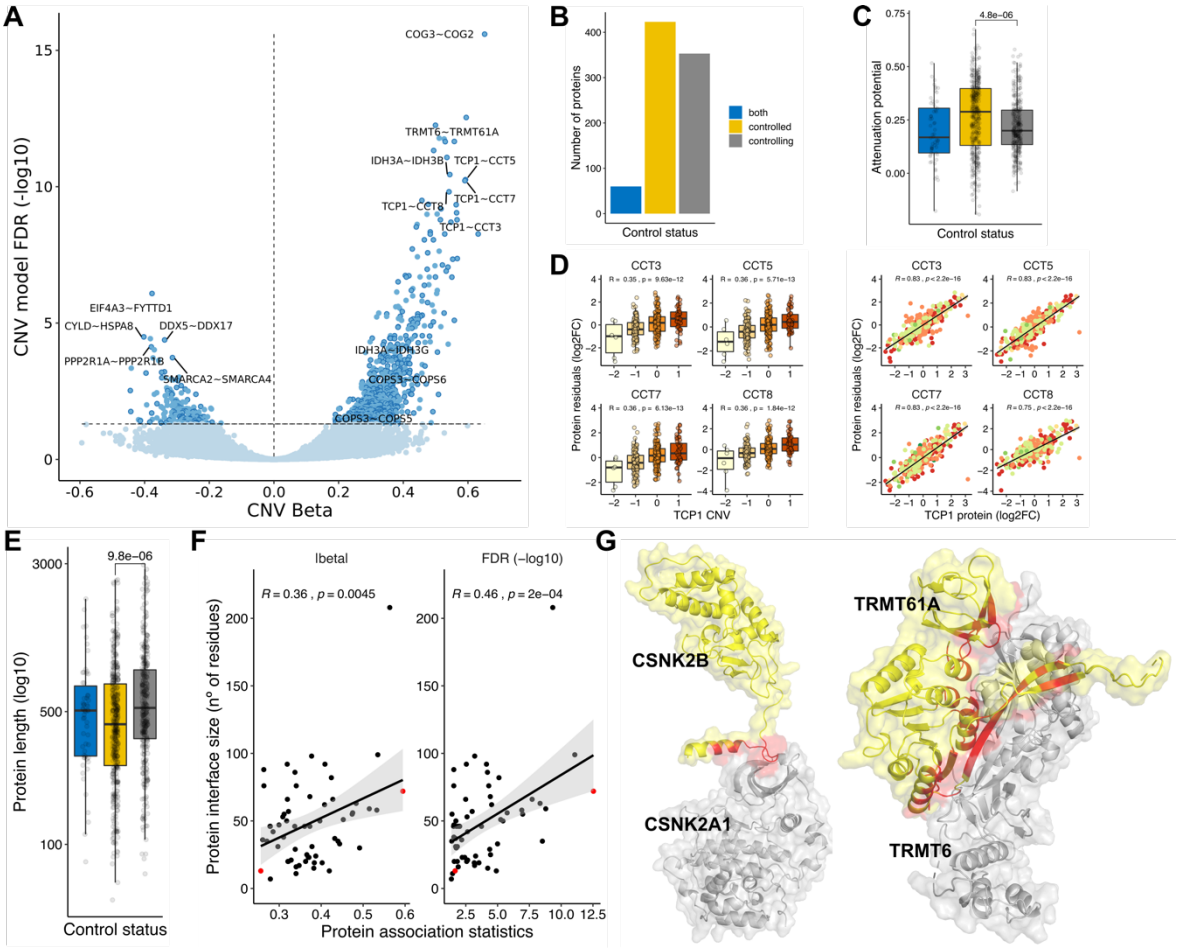


Figure 3.2. Physical protein associations. (A) Volcano plot of CNV beta (x-axis) and FDR (y-axis) for 411,591 protein pairs. Non-significant associations (*FDR > 5%*) are represented in light-blue, and significant associations (*FDR < 5%*) in dark blue. Associations also found to be significant (*FDR < 5%*) in the mRNA model and filtered by genomic co-localization are highlighted with a darker border (516). **(B)** Number of proteins by control status. **(C)** Distribution of attenuation potential by control status. **(D)** Examples of protein associations between TCP1 (*controlling* protein) and CCT3, CCT5, CCT7 and CCT8 (*controlled* proteins). The boxplots show the relation between the CNV changes of TCP1 and the protein residuals (log2FC) of the interacting partners. The scatter plots show the same relation with the protein abundance of TCP1. **(E)** Protein length (log10 of number of residues) by control status. **(F)** Scatter plots displaying the correlation between the protein association statistics (beta and FDR) with the protein interface size (number of residues at the protein interface, measured in the *controlled* protein). Each dot is a protein association. Two representative associations between CSNK2A1 - CSNK2B (small interface) and TRMT6 - TRMT61A (big interface) are denoted in red. **(G)** Representation of protein interactions between CSNK2A1 and CSNK2B and TRMT6 and TRMT61A. The *controlled* proteins are coloured in yellow (CSNK2B and TRMT61A) and the *controlling* proteins are coloured in grey (CSNK2A1 and TRMT6). The interface area is represented in red.

3.3.3. Identification of phosphorylation sites that may modulate protein complex assembly

The role of phosphorylation in modulating protein binding affinities has been well described (Betts et al. 2017; Nishi, Hashimoto, and Panchenko 2011; Beltrao et al. 2012). We reasoned we could use the multi-omics datasets to find protein interactions affected by phosphorylation, which in turn could impact complex assembly and protein degradation. Out of 368 samples with CNV, mRNA and protein measurements, 170 also had quantifications at the phosphosite level (**Figure 3.3A**). For this analysis, we used proteins and phosphosites measured in at least 50% of the 170 samples, corresponding to 8,546 proteins and 5,733 phosphosites.

Using the compendium of physical interactions (572,856 protein interactions), we tested whether the changes of a phosphosite Xp from protein X is associated with the protein levels of the interacting protein Y. As before, we used a linear regression model where the protein abundance of protein Y is predicted using the phosphosite levels of protein X (Xp), while taking into account the protein and CNV levels of protein X, the RNA of protein Y, and other covariates (**Methods**). Out of 315,772 phosphosite-protein pairs tested with this model, 11,672 associations were significant (*FDR < 5%*). To ensure the associations are directional, we overlapped these associations with the 516 protein-protein associations found with the CNV and mRNA models, identifying 32 overlapping associations (**Figure 3.3B; Table S3.4**). Our interpretation of these associations is that these phosphosites can regulate the protein interaction and thereby modulate the degradation of the complex subunits.

The 32 associations involve 28 phosphosites, and of these, 2 phosphosites are already known to regulate interactions (POLD3 S458 and MYH9 S1943) and an additional case (EIF3A S492) is not yet known to regulate protein interactions but is at the interface with other complex members (**Figure 3.3C**). EIF3A is predicted here to be a “rate-limiting” subunit of the eukaryotic initiation factor 3 complex and has been previously experimentally implicated in the control of protein levels of several of the other subunits (S. Wagner et al. 2014). One phosphosite of EIF3A (S492) showed a strong association with the protein levels of two other complex subunits (EIF3D and EIF3E). In line with this, we find that the copy-number of EIF3A correlates with the residual protein levels of EIF3D (i.e., after regressing out EIF3D mRNA levels) and that the phosphosite levels of EIF3A S492 correlates better with EIF3D protein residual than the EIF3A total protein levels (**Figure 3.3D**). We further confirmed that the residual phosphosite levels of EIF3A S492 phosphosite (once accounting for the protein abundance of EIF3A) is significantly correlated with the protein levels of EIF3D in both datasets analysed (**Supplementary figure 3.6**). However,

we found dataset specific differences in the correlation of the protein levels of EIF3A and EIF3D (**Supplementary figure 3.6**). Overall, these results suggest that EIF3A S492 may have an impact on protein complex assembly.

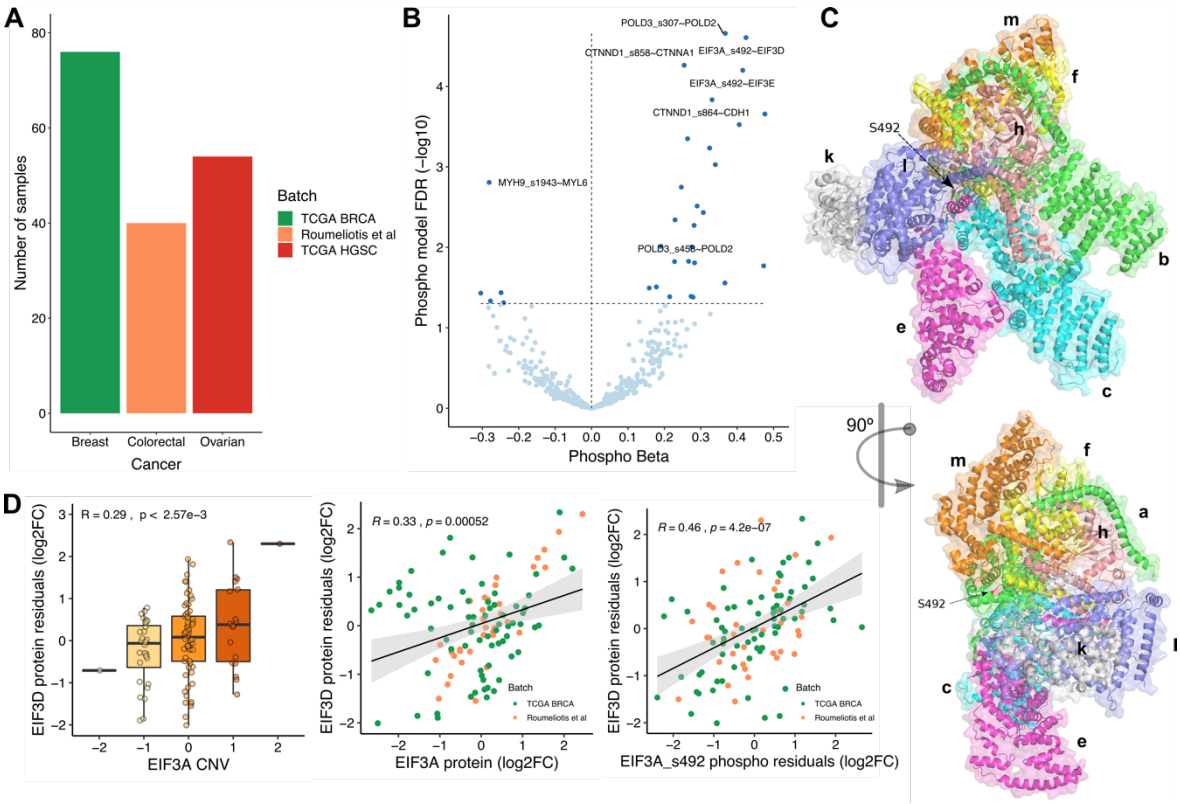


Figure 3.3. Identification of phosphorylation sites with a potential role in regulating protein interactions. (A) Number of samples with CNV, mRNA and phospho(protein) measurements, by cancer type/batch. **(B)** Volcano plot of phospho beta (x-axis) and FDR (y-axis). Each dot is a phosphosite-protein association, between a putative regulatory phosphosite Xp and a regulated protein Y. All associations (438) are significant in the CNV and mRNA models, between the putative regulatory protein X and the regulated protein Y. 32 associations (*FDR < 5%*) are also significant in the phospho model (dark blue). **(C)** Representation of EIF3 complex in two orientations. The arrow points to the phosphosite S492 (serine 492) at EIF3A subunit. **(D)** Significant association between EIF3A/EIF3A S492 and EIF3D. The boxplots show the agreement between the CNV changes of EIF3A and the protein residuals (log2FC) of EIF3D. The scatter plots show the same relation with the protein and phosphosite (S492) abundances of EIF3A.

3.3.4. Protein attenuation mechanisms found in cancer are observed in normal tissues

The study of the impact of CNVs in cancer proteomes indicates that up to ~40% of genes have copy-number changes that are buffered at the protein level. Such post-

transcriptional regulatory processes should not be specific to cancer, however, the extent that these effects are observed in normal cellular states is still largely unknown. To address this question, we analysed gene and protein expression datasets for normal tissues, made available by the Genotype-Tissue Expression (GTEx) and Human Protein Map (HPM) projects. In total, we collected expression measures for 5,239 proteins and genes, across 14 tissue types (**Methods**).

We tested if the post-transcriptional control dependent on protein interactions observed in cancer is present in normal tissues. For this, we asked if the protein abundance of *controlling-controlled* protein pairs will tend to correlate more strongly than other protein interaction pairs. Similarly, we expected that the correlation between the mRNA and protein levels of *controlled* subunits would tend to be weaker than for non-post-transcriptionally controlled proteins. We tested this using protein-protein interaction pairs measured in the tissue data with significant *controlling-controlled* relationships from cancer data (301 pairs), and all other 161,945 protein-protein interaction pairs (**Methods**). Reassuringly, we observed that the correlation of protein abundance across tissues increased for protein pairs with stronger association strength, for similar levels of mRNA-mRNA correlation values (Wilcoxon rank-sum test *P-value = 8.96e-4* between non-significant and significant pairs; *P-value = 8.25e-06* between non-significant and highly-significant pairs) (**Figure 3.4A**). Also, as predicted the protein to mRNA correlations across tissues of the *controlled* subunits decreases with the association strength (Wilcoxon rank-sum test *P-value = 0.022* between non-significant and significant pairs) (**Figure 3.4A**). We provide two examples for the protein interacting pairs ARCN1 and COPA and TRAPPC8 and TRAPPC11, where the mRNA levels of the *controlling* subunits (ARCN1 and TRAPPC8) appear to dictate the protein abundance of both proteins (**Figure 3.4B**). These results suggest that the protein associations identified in the cancer datasets can also be observed in normal tissues, at least in aggregate. Importantly, they demonstrate that cancer data can be a useful resource to study protein homeostasis in normal conditions.

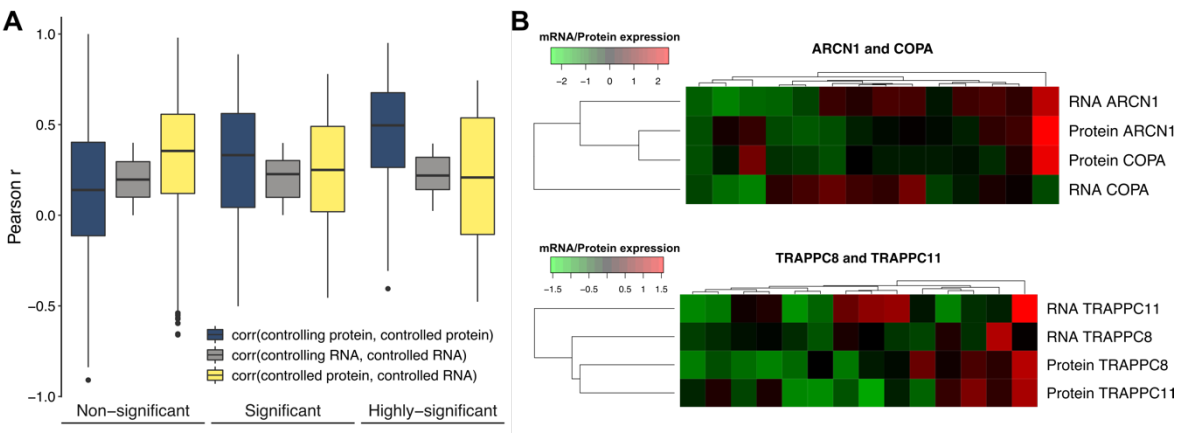


Figure 3.4. Evidence of interaction mediated control of protein abundances in normal tissues. (A) Pearson correlation coefficient between the protein of the *controlling* and *controlled* genes (blue); mRNA of the *controlling* and *controlled* genes (grey) and mRNA and protein abundance of the *controlled* gene (yellow); for the non-significant associations (*FDR > 5%*), significant associations (*1% < FDR < 5%*) and highly-significant associations (*FDR < 1%*). **(B)** Heatmap showing the agreement between the mRNA and protein expression profiles (rows) across tissues (columns) for two highly-significant associations: ARCN1 (*controlling*) ~ COPA (*controlled*) and TRAPPC8 (*controlling*) ~ TRAPPC11 (*controlled*).

3.3.5. Buffering of gene expression variation due to natural genetic variation

If mechanisms controlling the protein levels are consistent across cell types, then the attenuation models studied here could help to elucidate how natural variation may sometimes result in changes in mRNA but not protein and consequently phenotypic traits. Single nucleotide polymorphisms (SNPs) associated with gene expression via quantitative trait loci (QTL) analysis - known as expression QTLs (eQTLs) - should also tend to be attenuated at protein level potentially for the same genes as those found in cancer. To study this, we analysed if protein level CNV buffering could explain the probability of eQTLs to have phenotypic impact, i.e., in high linkage disequilibrium (LD) ($r^2 > 0.8$) with Genome Wide Association Studies (GWAS) variants (**Figure 3.5A; Methods;** on genes with significant CNV-mRNA Pearson's r > 0.3). To this end, we relied on *cis*-eQTLs reported in GTEx and compared the fraction of GWAS tagging eQTLs for different classes of protein attenuation (**Figure 3.5B; Methods**). We found that eQTLs corresponding to genes classified as highly attenuated have a lower fraction of GWAS tagging eQTLs, and that the difference between the degree of attenuation increases for eQTLs mapped in multiple tissues (**Figure 3.5B**).

Highly attenuated genes tend to be enriched in protein complexes and are likely essential to the cell, and therefore could have specific biases as to how eQTLs are linked to GWAS associated traits. To account for this potential bias, we replicated the analysis on members of protein complexes. Interestingly, this shows that the attenuation score has a higher impact on GWAS tagging probability for members of protein complexes, and more specifically for members of large protein complexes (>5 subunits) (**Supplementary figure 3.7**).

These results suggest that the CNV attenuation measured in cancer cells for protein abundance has direct application in the ranking of the potential impact of mRNA variation

on phenotypic differences, and support the idea that some of these attenuation mechanisms may take place in multiple tissues.

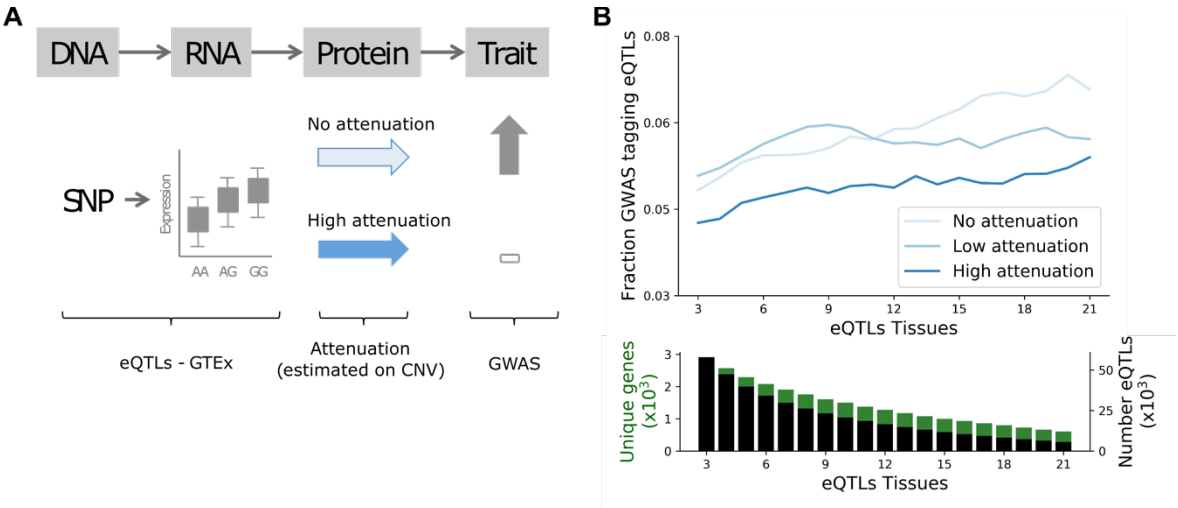


Figure 3.5. Protein attenuation reduces cis-eQTLs impact on phenotypic traits. (A) Illustration of the potential impact of protein attenuation on the eQTL associations with phenotypic changes. (**B**) Fraction of eQTLs associated with disease traits for the three classes of CNV attenuation: no (light blue), low (blue) and high (dark-blue) attenuation. The fractions are reported for increasing number of eQTL tissues, i.e., minimal number of tissues in which an eQTL was called (x-axis). Bottom panel shows the number of genes and eQTLs used in the top figure for cumulative stratification of eQTL tissues.

3.4. Discussion

The joint analysis of multi-omics datasets of cancer samples suggests that a very significant fraction of the proteome (up to 42%) is under post-transcriptional control. The set of genes with protein level buffering of CNVs is enriched in gene products belonging to large protein complexes. In addition, we found that the fraction of interface residues of a protein is a strong determinant of attenuation. Together with experiments on pulse chase degradation (McShane et al. 2016), aneuploidy (Dephoure et al. 2014; Stingele et al. 2012; Pavelka et al. 2010) and the impact of natural genetic variation on protein levels (Chick et al. 2016; Battle et al. 2015), these results implicate protein complex formation as an important factor in post-transcriptional control, most likely via a high degradation rate of unassembled subunits. We note that this mechanism of CNV buffering at the protein level may be possible with CNV amplifications and deletions. While in the former it would be manifested by an apparent increase in the degradation rate of free complex subunits, in the latter it would result from a decrease in the apparent degradation rate of free subunits.

However, it is likely that multiple mechanisms contribute to the post-transcriptional control measured in the cancer samples including, for example, the control of protein translation rates by microRNAs or RNA-binding proteins. The extent of post-transcriptional control that is explained by the different processes remains to be studied.

We observed that the fraction of residues at the interface correlates with the probability that a protein shows gene dosage attenuation. Similarly, the size of the interface correlates with the strength of association between pairs of physical interactions in which one subunit appears to control the abundance level of the interaction partner. The size of the interface typically correlates with increasing binding affinity between proteins as well as larger amounts of hydrophobic residues that are exposed in the absence of interactions. We speculate that either of these consequences could play a role in the attenuation. In particular, larger fractions of hydrophobic regions could increase the propensity to form aggregates and, in some cases, hydrophobic regions are known to be recognized for degradation (Xu, Anderson, and Ye 2016). This could represent a general mechanism for recognition of unassembled complex subunits. The structural analysis performed here is limited by the current lack of coverage for structures of protein complexes. In the future, additional structures may allow us to study in more detail the interface features that are important for the attenuation mechanism.

We have used data from cancer samples to identify the attenuated proteins and physical interactions with rate-limiting subunits. We find that most of the *controlling-controlled* protein-protein associations we predict have a positive relationship. Given the working model that these are explained by protein complex formation, the negative associations could be explained by cases of mutually exclusive complex membership. The fact that few predicted associations are negative is consistent with the idea that most complex members are not mutually exclusive.

It is still unclear if the same proteins and interactions will have the same post-transcriptional control in other systems and/or species. When studying expression variation in normal tissues and the association of eQTLs with phenotypes we observed that, in aggregate, the same proteins and interactions show signals consistent with post-transcriptional buffering of mRNA expression variation. Of note, we find that eQTLs are less likely to be linked to phenotypes in highly attenuated proteins. This is in line with studies of mRNA and protein QTLs in human induced pluripotent stem cell lines (iPSCs), showing that genetic variation driving mRNA changes is more likely to be associated with phenotype differences when they are observed at the protein level (Mirauta et al. 2020). These findings highlight the importance of studying the degree of conservation of these post-transcriptional processes in different tissues and systems in the context of human genetics and disease.

3.5. Methods

Multi-omics data collection

Proteomics and phosphoproteomics quantifications at the protein/phosphosite level from TCGA cancer patients were obtained from the CPTAC data portal (proteomics.cancer.gov/data-portal), for breast cancer (BRCA) (Mertins et al. 2016), colorectal cancer (COREAD) (B. Zhang et al. 2014) and ovarian cancer (HGSC) (H. Zhang et al. 2016). The same data from cancer cell lines were downloaded for COREAD cell lines (Roumeliotis et al. 2017) and for BRCA cell lines (Lapek et al. 2017; Lawrence et al. 2015). Gene-level RNA-seq raw counts were acquired from GEO (GSE62944) (Rahman et al. 2015) for TCGA samples and from the CCLE data portal (portals.broadinstitute.org/ccle/data) (Stransky et al. 2015; Barretina et al. 2012) for cancer cell lines. Gene copy-number profiles in this study were represented using discretized GISTIC 2.0 scores as described here (Mermel et al. 2011; Beroukhim et al. 2007). Briefly, these discrete variables can be -2 (strong copy-number loss, likely a homozygous deletion); -1 (shallow deletion, likely a heterozygous deletion); 0 (diploid); 1 (low-level gain of copy-number, generally broad amplifications) and 2 (high-level increase in copy-number, often focal amplification). CNV GISTIC 2.0 levels were compiled from the firebrowse (firebrowse.org/) data portal (accession date 15/01/2018) for TCGA samples and from the CCLE data portal for cancer cell lines (accession date 14/02/2017).

Data pre-processing and normalisation

The label-free protein quantifications (precursor areas) for COREAD CPTAC samples (B. Zhang et al. 2014) were first normalized by sample, where summed peak areas for the same protein were divided by the total summed area for the observed sample proteome. Relative protein abundances were then calculated by dividing each protein area over the median area across samples, and then log2 transformed. Protein and phosphosite intensities for COREAD cell lines (Roumeliotis et al. 2017) were divided by 100 and transformed to log2. For BRCA cell lines (Lapek et al. 2017) protein log2 fold-changes were calculated by subtracting the median intensities across the samples. Similarly, the label-free protein intensities (peak areas) for BRCA cell lines from (Lawrence et al. 2015) were

converted into relative abundances by calculating the log2 ratio of protein intensities over the median intensities across samples. Sample replicates of protein and phosphoprotein were combined by averaging the values for each protein and phosphosite, respectively. Phosphopeptides intensities mapping to the same phosphosite were combined by calculating the median phosphosite intensity per sample. In the cancer cell lines, genes with multiple isoforms were filtered by selecting the protein isoform with highest median expression across samples. Proteomics and phosphoproteomics distributions across cancer samples and cell lines were quantile normalized to ensure comparable distributions, using the *normalizeQuantiles* function from Limma R package (Ritchie et al. 2015). In total, 13,569 proteins across 436 samples (340 cancer samples and 96 cell lines) and 79,824 phosphosites across 195 samples (145 cancer samples and 50 cell lines) were assembled in this study. Given the sparseness of the phospho(protein) data, for the subsequent analyses we only selected proteins measured in at least 25% of the 368 samples with protein, mRNA and CNV measurements, and the phosphosites measured in at least 50% of the 170 samples with also phosphorylation data, comprising 8,124 proteins and 5,733 phosphosites. The phospho(protein) and mRNA data were then standardized using the z-score transformation.

At the RNA-seq level, lowly expressed genes were removed by filtering out genes with mean counts-per-million (CPM) lower than 1 across samples. After raw counts normalization by the trimmed-mean of M-values method (Robinson and Oshlack 2010) using the edgeR R package (Robinson, McCarthy, and Smyth 2009), the log2-CPM values were extracted from the *voom* (Law et al. 2014) function in Limma. After merging the CPTAC samples with the CCLE cell lines, the final RNA-seq dataset comprised 13,228 genes with measurements across 370 samples (296 cancer samples and 74 cell lines).

At the CNV level, after compiling the GISTIC 2.0 thresholded data, 19,023 genes were found to have CNV measurements across 412 samples (337 cancer samples and 75 cell lines).

Potential confounding factors revealed by PCA analysis (**Supplementary figure 3.1A, Supplementary figure 3.2A**) were regressed-out using a multiple linear regression model. This model was implemented with the protein or mRNA abundance of a given gene as dependent variable and the potential confounding factors, i.e., cancer type, experimental batch, proteomics technology, age and gender as independent variables. The residuals from the linear model were the protein and mRNA variation not driven by the confounding effects, as the second PCA demonstrated (**Supplementary figure 3.1B, Supplementary figure 3.2B**).

Analysis of protein attenuation

The strategy in (Gonçalves et al. 2017) was used to evaluate the impact of CNVs at the genome level on cancer proteomes. For each gene, the Pearson correlation coefficients between the CNV and mRNA and the CNV and protein were calculated, and an attenuation measure devised as follows:

Equation 3.1: Attenuation potential$_i$ = corr(CNV$_i$, mRNA$_i$) - corr(CNV$_i$, Protein$_i$), i $\in$ Protein

where corr represents the Pearson correlation coefficient and Protein represents 8,124 genes for which CNV, mRNA and protein quantifications across 368 samples where available. After calculating the attenuation potentials, a GMM model with 4 mixture components was used to cluster the genes in four different groups. Group 1 had 19 genes with a negative attenuation potential, due to the higher correlation between the CNV and Protein than with the CNV and mRNA. These genes, which were not attenuated at the protein level, were included with the remaining non-attenuated genes in group 2, comprising 4,689 genes. Groups 3 and 4 contained the lowly-attenuated and highly-attenuated genes, with 2,578 and 857 genes, respectively. The GMM was implemented using the *Mclust* function from the mclust R package (Scrucca et al. 2016).

The enrichment of CORUM complexes was calculated with a hypergeometric model, using the *enrichr* function from the clusterProfiler R package (G. Yu et al. 2012). Only CORUM complexes with a Jaccard index lower than 0.9 and with more than 5 proteins were used. The comparison of ubiquitination site fold-changes across protein attenuation levels was done using protein ubiquitination data obtained with three proteasome inhibitors: MG-132, epoxomicin and bortezomib (Higgins et al. 2015; Udeshi et al. 2013; W. Kim et al. 2011; S. A. Wagner et al. 2011).

Compendium of physical protein interactions

In order to build a compendium of physical protein interactions, we downloaded a data set of protein-protein interactions from BioGRID version 3.4.157 (C. Stark et al. 2006) (accession date 30/01/2018). We only selected protein interactions occurring in humans and captured with physical experimental systems. Interactions captured with Affinity Capture-RNA and Protein-RNA were excluded in order to guarantee that our dataset

contained only interactions observed at the protein level. After excluding protein homodimers, 524,148 protein interactions (262,074 unique) were compiled with BioGRID. A list of protein interactions was also built using a set of protein complexes from the CORUM database (Giurgiu et al. 2019) (accession date 29/05/2018). The rationale was that protein partners from the same protein complex interact physically at least once. Using a set of 1,787 protein complexes and excluding protein homodimers, we assembled 74,712 (37,356 unique) physical protein interactions. A small number of 890 endoplasmic reticulum-related interactions were additionally curated from the literature. In total, 572,856 (286,428 unique) protein physical interactions were compiled.

Linear modelling to identify protein and phospho-protein associations

Protein associations:

For a given protein physical interaction pair X and Y, it was tested whether protein X can control the protein levels of Y through protein-protein interactions, potentially constraining the degradation rate of Y. For each interacting pair two nested linear models were fitted. The first model (null) was used to predict the protein levels of Y (Py) using its mRNA (Ty) and a set of other covariates, i.e., cancer type, experimental batch, proteomics technology, patient age and gender (**Equation 3.2**). In a second linear model (alternative), the CNV levels of X (Gx) were added as a predictor variable (**Equation 3.3**). A likelihood ratio test (LRT) (**Equation 3.4**) was then applied in order to test whether the second model increases the goodness of fit of the first model in predicting Py.

Equation 3.2: Null model: $P_y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 T_y + [\boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\beta}_5, \boldsymbol{\beta}_6] + \varepsilon$

$\boldsymbol{\beta}_0$ represents the intercept, $\boldsymbol{\beta}_1$ the regression coefficient (effect size) for the mRNA of Y, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_4$, $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_6$ the regression coefficients for the covariates cancer type, experimental batch, proteomics technology, age and gender, respectively. ε is the noise term.

Equation 3.3: Alternative model: $P_y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 T_y + [\boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\beta}_5, \boldsymbol{\beta}_6] + \boldsymbol{\beta}_7 G_x + \varepsilon$

β_7 is the regression coefficient for the CNV (G_x) of protein X. An LRT was used to assess the significance of the association:

Equation 3.4: $LRT = 2 \times [\log_e Lik(Alternative) - \log_e Lik(Null)]$

$\log_e Lik$ corresponds to the log likelihood of the alternative and null models. *P-values* were then calculated using the LRT statistic over a chi-squared distribution and adjusted for FDR using the Benjamin-Hochberg method. This model was applied for a given protein association pair X and Y if: $X \in CNV \wedge Y \in Protein \wedge Y \in mRNA$, where CNV, Protein, and mRNA are the sets of genes detected with the respective assays.

A total of 411,591 protein pairs followed these criteria and were tested across 368 tumor samples. The same analysis was performed with the mRNA, instead of the CNV, of protein X for 392,128 protein pairs. To avoid spurious protein associations that might occur due to the genomic co-localization of the *controlling* proteins, the top-ranked association was selected using the Borda ranking method. This was done systematically for all cases where multiple *controlling* proteins in the same chromosome were associated with the same *controlled* protein. More than one *controlling* protein in the same chromosome for the same *controlled* protein was allowed if their CNV profile Pearson correlation was lower than 0.5.

The linear models were implemented using the *lm* R function. The LRT tests with associated statistics were calculated using the *lrtest* function from the lmtest R package. The Borda ranking method was implemented using the *Borda* function from the TopKLists R package (Schimek et al. 2015).

Phospho-protein associations:

For a given protein pair X and Y, it was tested whether a phosphosite Xp from protein X can be associated with changes in the protein abundance of protein Y. A similar model to the before linear regression models and LRT tests was used. For each phosphosite-protein interaction, a first null model was fitted to predict the protein levels of Y (Py) using its mRNA (Ty), the CNV and protein levels of protein X (Gx and Px), and the covariates experimental batch, patient age and gender (**Equation 3.5**). In a second alternative linear model, the phosphosite Xp (Phox) of protein X was added as a predictor variable (**Equation 3.6**). The models were then compared using an LRT as in **Equation 3.4**.

Equation 3.5: Null model: $P_y = \beta_0 + \beta_1 T_y + \beta_2 G_x + \beta_3 P_x + [\beta_4, \beta_5, \beta_6] + \varepsilon$

where β_0 represents the intercept, β_1 the coefficient of the mRNA of Y, β_2 and β_3 the regression coefficients for the CNV and Protein of X, respectively, and β_4, β_5 and β_6 the regression coefficients for the covariates experimental batch, age and gender, respectively. ε is the noise term.

Equation 3.6: Alternative model: $P_y = \beta_0 + \beta_1 T_y + \beta_2 G_x + \beta_3 P_x + [\beta_4, \beta_5, \beta_6] + \beta_7 Pho_x + \varepsilon$

where β_7 is the regression coefficient for the phosphosite Xp of protein X. This model was applied for a given phosphosite-protein association pair Xp and Y if: $Xp \in Phospho \wedge X \in Protein \wedge X \in CNV \wedge Y \in Protein \wedge Y \in mRNA$, where Phospho, Protein, CNV and mRNA are the sets of genes detected with the respective assays. A total of 315,772 phosphosite-protein pairs followed these criteria and were tested with this model across 170 tumor samples.

Structural analysis

Protein interface sizes were calculated using an in-house pipeline (int3dInterfaces, github.com/evocellnet/int3dInterfaces) that extracts protein interfaces from Interactome3D structures (Mosca, Céol, and Aloy 2013). For each protein interaction structure in Interactome3D, this pipeline uses NACCESS (bioinf.manchester.ac.uk/naccess) to calculate the solvent accessibility of the bound and unbound monomers. Every residue changing its relative solvent accessibility is considered to form part of the interface. From the 11,530 human protein interaction structures analysed with this pipeline, structures of protein homodimers or structures with less than 100 amino acids were removed. Also, structures with chain lengths bigger than the respective UniProt protein lengths and with the same chain length for each partner were removed. After applying these filters, 3,082 structures with 6,147 protein interactions were used in the subsequent analyses.

For the 1,470 proteins which contained both information about CNV attenuation and interface size, the percentage of residues in protein interfaces was calculated as the ratio of the number of unique residues in interfaces over the protein size. For 60 significant protein association pairs represented in the structural data, the relation between the protein interface size with the regression CNV coefficient and FDR was assessed using the Pearson correlation coefficient. For each pair, the protein interface size was calculated in the *controlling* and *controlled* proteins. The protein sizes (number of residues) were obtained from UniProt for 20,349 proteins (accession date 19/06/2018).

The percentage of area inside the complex for the protein subunits from the COP9 signalosome was calculated using FreeSASA (Mitternacht 2016). For each protein subunit, this percentage corresponded to the difference between the solvent accessible surface area (SASA) outside and inside complex over the SASA outside complex. The SASA was calculated in units of squared Ångström ($Å^2$).

Analysis of gene essentiality using CRISPR-Cas9 screenings

Gene essentiality data obtained with CRISPR-Cas9 screenings (Meyers et al. 2017) was downloaded from Project Achilles data portal (depmap.org/portal/achilles/) (accession date 31/10/2017). These data contain gene-dependency levels adjusted for copy-number specific effects for 17,670 genes across 341 cancer cell lines. Genes with an essentially score lower than -1×SD (the standard deviation for the entire data set corresponds to 0.3) in more than 5% of the cell lines were considered essential and used in the remaining analysis (5,532 genes). The median gene essentiality was calculated for 3,548 genes with attenuation and essentiality data across the 341 cancer cell lines.

Pairwise correlation of protein association pairs using normal tissue data

Gene and protein expression data for normal human tissues were obtained from the GTEx (gtexportal.org/) (Aguet et al. 2017) and HPM (humanproteomemap.org/) (M. S. Kim et al. 2014) data portals. The gene expression was obtained in the format of RNA-seq median RPKM for 56,238 genes across 53 tissues. The protein expression was downloaded as averaged label-free spectral counts for 17,294 genes across 30 tissues. For the protein expression data, it was selected 9,156 genes in common with the HPM data available in Expression Atlas (Petryszak et al. 2014). The 14 tissues common to the GTEx and the HPM used in the remaining analysis were: frontal cortex, spinal cord, liver, ovary, testis, lung, adrenal gland, pancreas, kidney, urinary bladder, prostate gland, heart, esophagus and colon. The gene expression in the last three tissues was averaged in GTEx, between heart atrial appendage and left ventricle; between esophagus gastroesophageal junction, mucosa and muscularis; and between colon sigmoid and transverse. The protein and gene expression data was then filtered to only include genes and proteins expressed in at least 10 of 14 tissues, resulting in 5,239 genes consistently expressed at the gene and protein

level. The RNA and protein measurements were then standardized to z-scores and quantile normalized.

Having assembled the gene and protein expression datasets for normal tissues, pairwise Pearson correlation coefficients were calculated between the protein of the *controlling* and *controlled* genes, mRNA of the *controlling* and *controlled* genes, and mRNA and protein of the *controlled* gene. The Pearson correlations were calculated for 91 highly-significant associations ($FDR < 0.01$), 210 significant associations ($0.01 \leq FDR < 0.05$) and 161,945 non-significant associations at the CNV and mRNA level ($FDR \geq 0.05$). In order to assure that the increase in protein-protein correlations were not simply due to an increase in mRNA-mRNA correlations, we selected the protein pairs with mRNA Pearson's correlation coefficient between 0 and 0.4, corresponding to 57,145 pairs (30 highly-significant, 69 significant and 57,046 non-significant).

Analysis of the impact of CNV attenuation on the eQTL association to disease traits

Following the approach in HipSci proteomics (Mirauta et al. 2020), we considered a stringent set of 21,601 associations from the NHGRI-EBI GWAS catalog (download on 10 April 2018; converted to hg19) for analysis. We considered eQTLs reported from GTEx in 35 tissues (excluding brain), computed the number of tissues having the same slope sign, i.e., direction of effect size, and discarded those with consistent slope in less than 3 tissues.

We defined proxy variants of each *cis*-eQTL as variants in high LD ($r^2 > 0.8$; based on the UK10K European reference panel) within the same *cis* window. Next, we grouped the eQTLs in high LD blocks ($r^2 > 0.8$), excluded from this analysis 247 genes having each more than 100 eQTL blocks, and obtained a final set of 66,197 eQTL blocks corresponding to 2,953 genes and 441,194 eQTL-gene associations. We then defined these blocks as GWAS-tagging if for at least one eQTL in the block at least one LD proxy variant was annotated in the NHGRI-EBI GWAS catalog. Finally, we reported the fraction of GWAS-tagging eQTLs stratified by the attenuation level of the corresponding *cis* genes. To assess the robustness of this analysis and to study the effects on GWAS tagging probability of eQTL recurrence across tissues, we calculated the number of tissues in which an eQTL was called with the same slope and reported the results by stratifying the eQTLs by increasing number of tissues.

We relied on core human protein complexes from CORUM to identify the gene complex membership status and segregated those that are annotated in at least one large

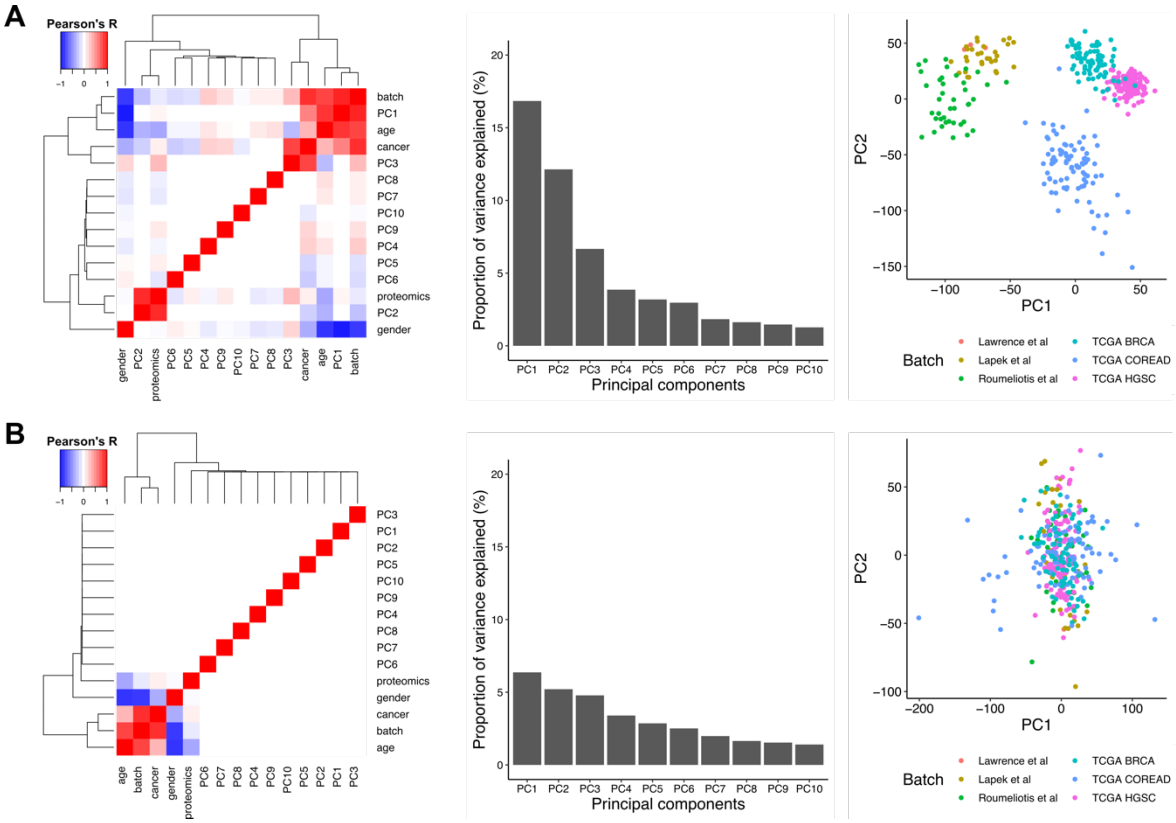
complex (>5 subunits). Out of the genes with eQTL evidence and with annotation scores, 961 were annotated in CORUM and 576 were members of large complexes.

Code availability

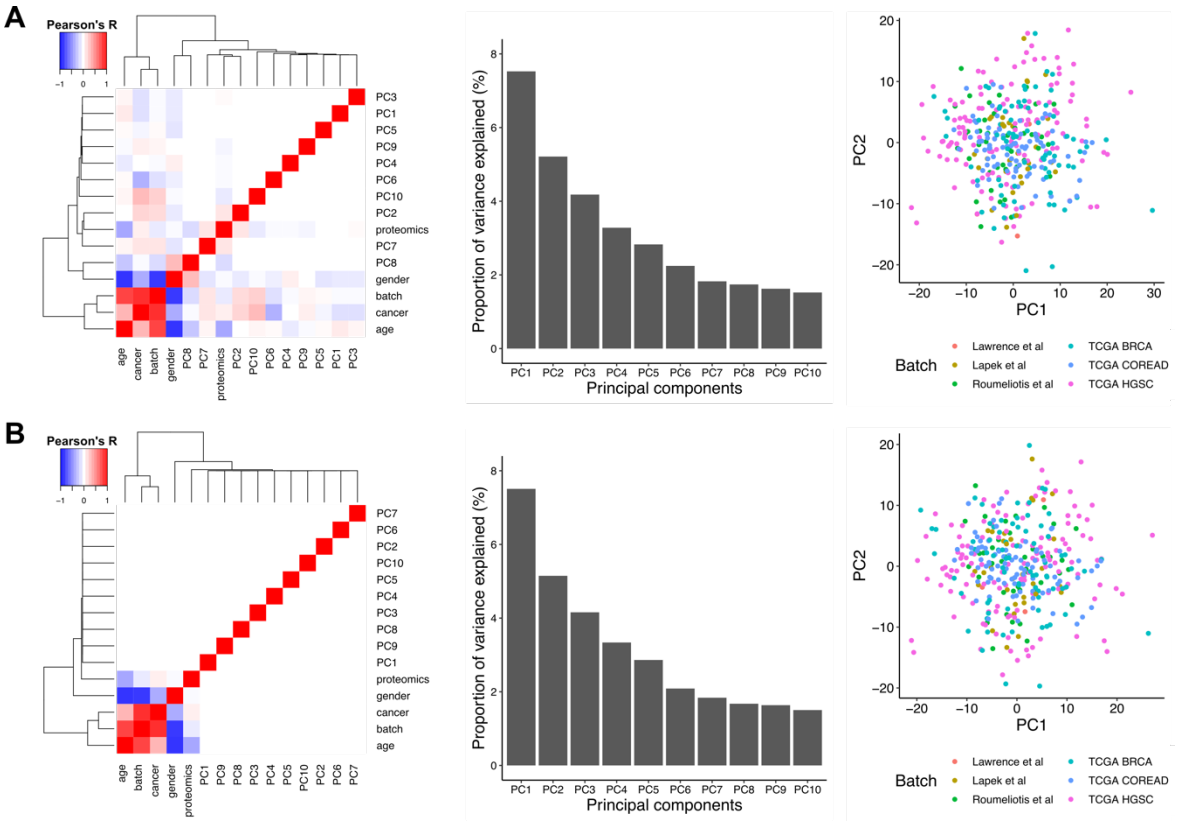
All the code is available under a GNU General Public License V3 in a GitHub project, at the following url: github.com/abelfsousa/cnv_buffering.

3.6. Supplementary materials

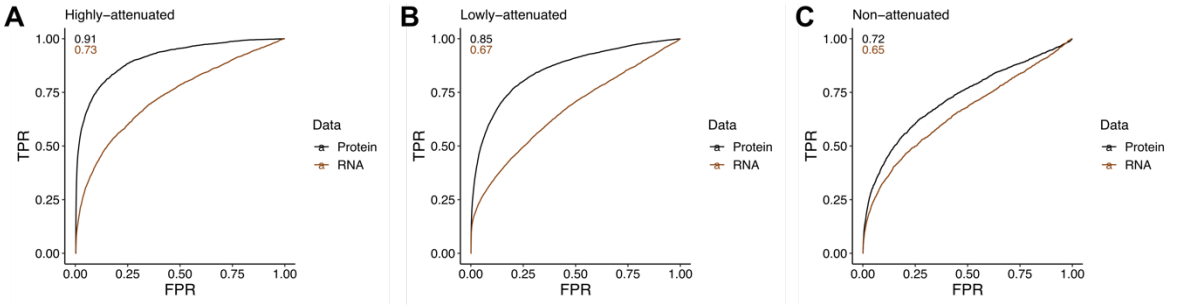
3.6.1. Figures



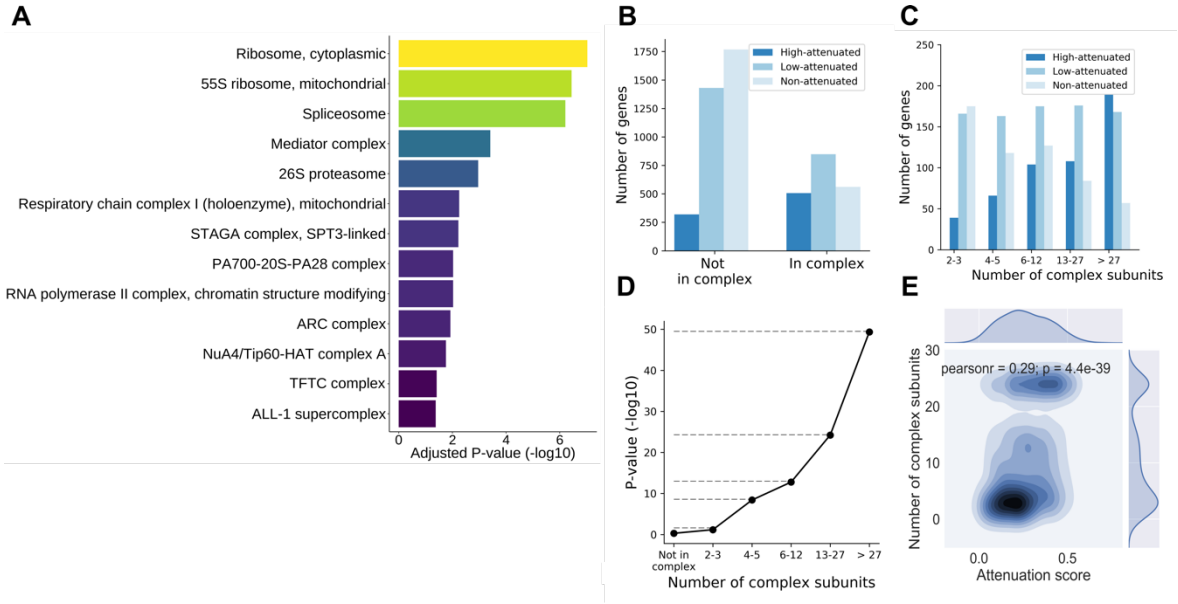
Supplementary figure 3.1. Confounding effects regressed-out from transcriptomics data. (A) Pearson correlation coefficient of the first 10 principal components (PCs) with the potential confounding effects before normalization. **(B)** Pearson correlation coefficient of the first 10 principal components (PCs) with the potential confounding effects after normalization.



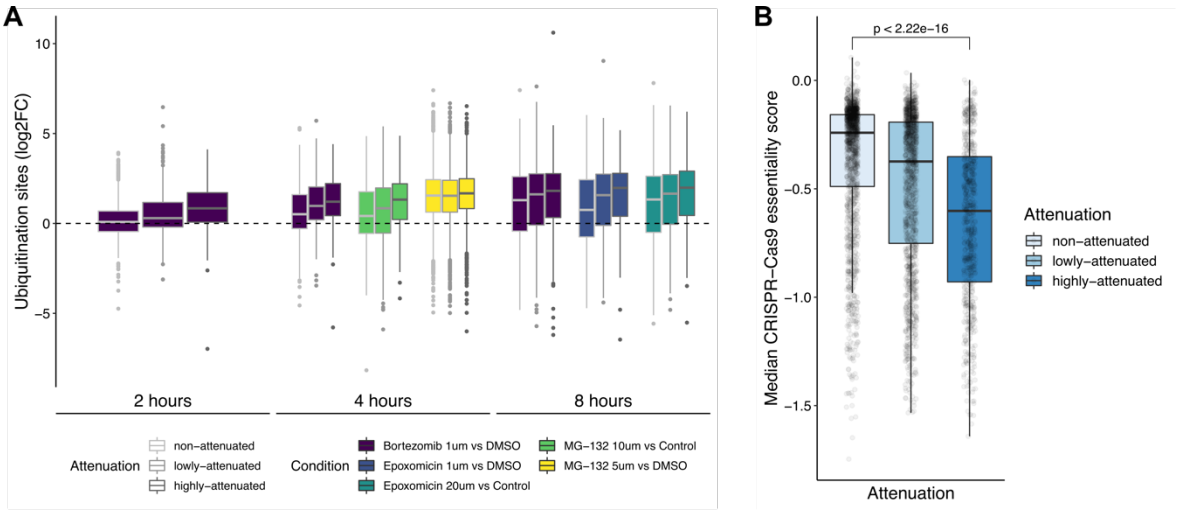
Supplementary figure 3.2. Confounding effects regressed-out from proteomics data. **(A)** Pearson correlation coefficient of the first 10 principal components (PCs) with the potential confounding effects before normalization. **(B)** Pearson correlation coefficient of the first 10 principal components (PCs) with the potential confounding effects after normalization.



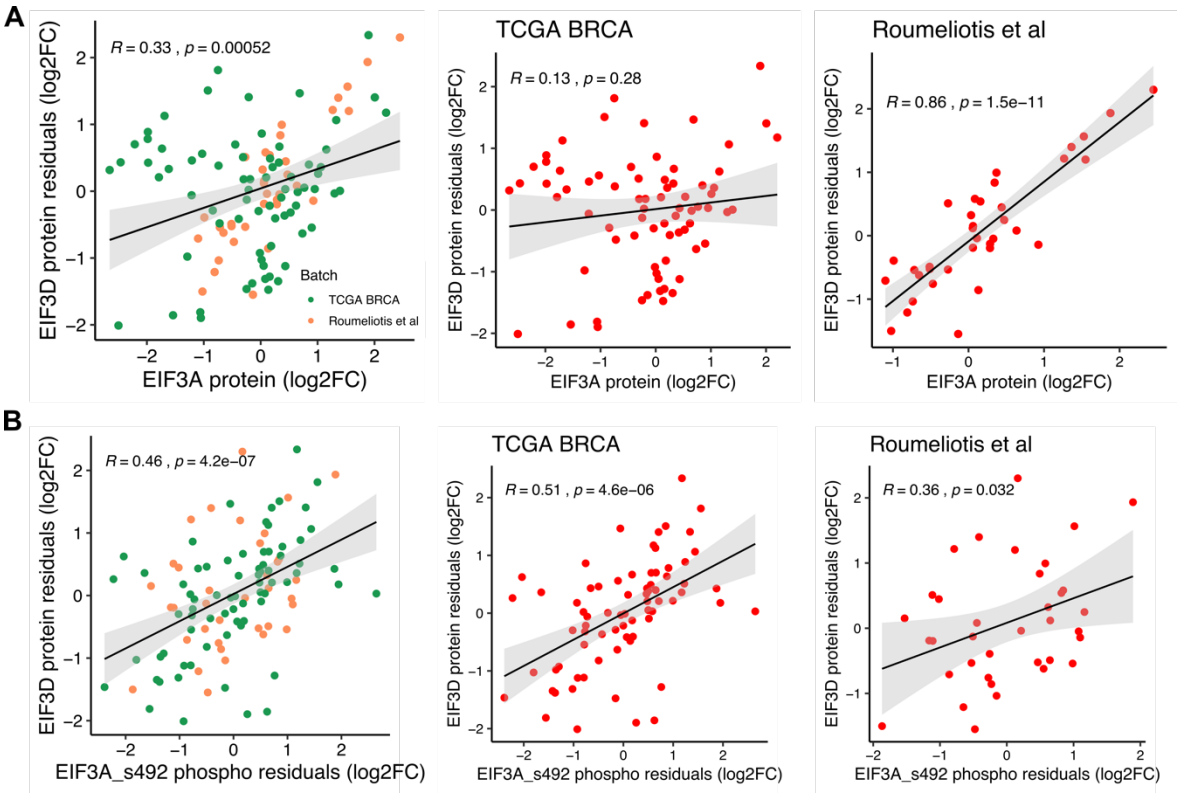
Supplementary figure 3.3. ROC curve analysis for the prediction of true protein-protein interacting pairs. Correlation of protein abundance and gene expression was used as a predictor of CORUM protein pairs, among the highly-attenuated **(A)** lowly-attenuated **(B)** and non-attenuated **(C)** protein pairs. The x-axis represents the false-positive rate (FPR) and the y-axis the true-positive rate (TPR). The AUC of each curve is indicated.



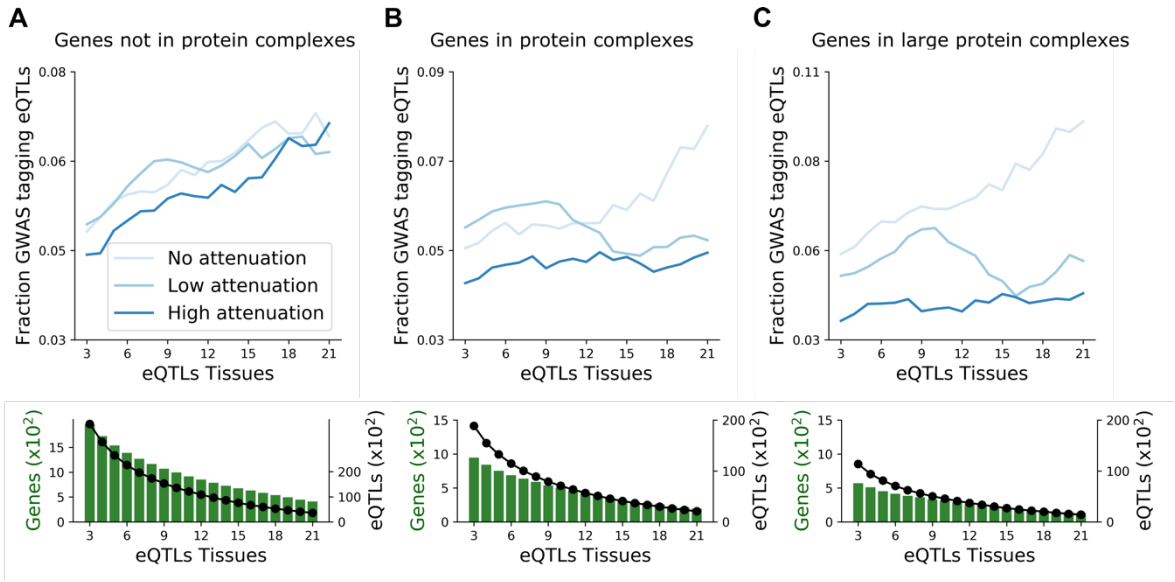
Supplementary figure 3.4. Relationship between CNV attenuation in proteins and protein complex membership. (A) List of top protein complexes ordered by enrichment in attenuated genes. X-axis shows P-values derived with an hypergeometric test (Benjamini-Hochberg multiple testing correction). We discarded from the list of CORUM complexes those with a Jaccard index higher than 0.9 with any other complex and those with 5 proteins or less **(B)** Number of genes for each attenuation class by complex membership status (CORUM). **(C)** Number of genes stratified by the maximum number of subunits of any protein complex incorporating the genes. **(D)** Enrichment of attenuated proteins in members of protein complexes stratified by complex size, i.e., number of subunits. Shown are the P-values derived with a Fisher's exact test (*alternative* = *"greater"*). **(E)** Relationship between the number of subunits in a protein complex and the protein complex member CNV attenuation. For each member of a complex we juxtapose the attenuation score (x-axis) and the maximum number of subunits of any complex this protein is part of (y-axis).



Supplementary figure 3.5. Attenuated proteins show faster increase in protein ubiquitination after proteasome inhibition and higher gene essentiality. (A) Ubiquitination sites fold-changes (y-axis) across protein attenuation levels (x-axis) after proteasome inhibition with three inhibitors: Bortezomib, Epoxomicin and MG-132. **(B)** Median gene essentiality measured in CRISPR-Cas9 screenings (y-axis), across 341 cancer cell lines, by attenuation level (x-axis).



Supplementary figure 3.6. Correlation of EIF3A protein and EIF3A S492 phosphosite with EIF3D protein. The scatterplots show the Pearson's correlation coefficient between EIF3A protein log2 fold-changes **(A)** and EIF3A S492 phosphosite log2 fold-changes **(B)** with EIF3D protein residuals (after regressing-out the mRNA and possible confounding factors from protein expression). The EIF3A protein abundance and confounding factors were also regressed-out from EIF3A S492 phosphorylation levels. The correlations are shown with all samples and by dataset (TCGA breast cancer samples (BRCA) and colorectal cancer cell lines from *Roumeliotis et al.*).



Supplementary figure 3.7. Impact of the CNV attenuation at protein level on the eQTL association with disease traits, stratified by the protein complex membership status.

Identical analysis as in **Figure 5 (B)** is performed for genes with no **(A)** and with existing annotation in CORUM protein complexes **(B)** and **(C). (C)** for genes members of protein complexes with at least 5 subunits. Bottom panels: the number of genes and eQTLs considered for the analysis shown in the top panels.

3.6.2. Tables

Supplementary table 3.1. CNV, mRNA and protein measurements across cancer samples for 8,124 genes. The CNV data is represented as discretized GISTIC 2.0 scores (**Methods**). The mRNA and protein measurements are represented as z-scores, with potential confounding factors regressed-out using a multiple linear regression model (**Methods**). The cancer type (breast, colorectal and ovarian), experimental batch (TCGA BRCA, TCGA COREAD, TCGA HGSC, *Lawrence et al.*, *Roumeliotis et al.* and *Lapek et al.*) and proteomics type (TMT and label-free) of each sample are also included as columns.

Supplementary table 3.2. 8,124 genes stratified by attenuation level. The table includes the Pearson correlation coefficients between the CNV and mRNA and the CNV and protein, respective *P-values* and attenuation potentials.

Supplementary table 3.3. 516 protein-protein associations significant in the CNV and mRNA models (*FDR < 5%*). For each association, the table includes the *controlling* and *controlled* proteins and the effect size (beta) and FDR from both models.

Supplementary table 3.4. 32 significant phospho-protein associations (*FDR < 5%*). The table includes the *controlling* protein/phosphosite, the *controlled* protein, and the effect size (beta) and FDR from the phospho model. All associations are also significant in the CNV and RNA models, between the putative regulatory and regulated proteins.

All supplementary tables can be consulted using the following DOI:
doi.org/10.1074/mcp.RA118.001280

4. Pan-Cancer Landscape of Protein Activities Identifies Drivers of Signalling Dysregulation and Patient Survival

This chapter includes published material from the following article:
Abel Sousa, Aurelien Dugourd, Danish Memon, Borgthor Petursson, Evangelia Petsalaki, Julio Saez-Rodriguez and Pedro Beltrão. Pan-Cancer Landscape of Protein Activities Identifies Drivers of Signalling Dysregulation and Patient Survival. BioRxiv, 9 June 2021.

4.1. Abstract

Genetic alterations in cancer cells trigger oncogenic transformation, a process largely mediated by the dysregulation of kinase and transcription factor (TF) activities. While the mutational profiles of thousands of tumours have been extensively characterized, the measurements of protein activities have been technically limited until recently. We compiled public data of matched genomics and (phospho)proteomics measurements for 1,110 tumours and 77 cell lines that we used to estimate activity changes in 218 kinases and 292 TFs. Kinase activities are, on average, not strongly determined by protein abundance but rather by their phosphorylation state while the reverse is more common for TFs. Co-regulation of kinase and TF activities reflects previously known regulatory relationships and allows us to dissect genetic drivers of signalling changes in cancer. Loss-of-function mutation is not often associated with dysregulation of downstream targets, suggesting frequent compensatory mechanisms. Finally, we identified the activities most differentially regulated in cancer subtypes and showed how these can be linked to differences in patient survival. Our results provide broad insights into dysregulation of protein activities in cancer and their contribution to disease severity.

4.2. Introduction

Cancer is a highly heterogeneous disease that is generally caused by the acquisition of somatic genomic alterations, including single nucleotide variants (SNVs), gene copy-number variations (CNVs) and large chromosomal rearrangements (Pleasance et al. 2010; Beroukhim et al. 2010; Campbell et al. 2020). The Cancer Genome Atlas (TCGA) has led to an in-depth characterization of the genomic alterations of more than 10,000 tumours from 33 cancer types (Hoadley et al. 2018; Ding et al. 2018). However, mutations in key driver genes are just the first steps of a cascade of events that culminate in tumour formation and cancer. These mutations generate the genetic diversity that promotes the acquisition of multiple cancer hallmarks, including chronic proliferation, resistance to cell death and tissue invasion and metastasis (Hanahan and Weinberg 2011). An understanding of the molecular mechanisms that underpin the development of cancer is critical in order to study cancer biology and to develop therapies.

While somatic alterations and gene expression changes across tumours have been extensively studied, key driver genomic changes in cancer are thought to result in changes

in cell signalling including the dysregulation of protein kinases and transcription factors (Yaffe 2019; Blume-Jensen and Hunter 2001). As an example, about 40% of melanomas contain the V600E activating mutation in the BRAF kinase, resulting in constitutive signalling through the Raf to mitogen-activated protein kinase (MAPK) pathway and increased cellular proliferation (Davies and Samuels 2010). Likewise, aberrant transcription factors (TFs) activities are a key feature of cancer cells (Garcia-Alonso et al. 2018). TFs are commonly dysregulated due to genomic alterations in their sequences or in upstream signalling regulatory proteins (Oliner et al. 1992; Ohh et al. 2000). Because of their role as signalling effectors, aberrant kinase signalling may dysregulate the activities of TFs and alter the expression of their target genes. Consequently, kinases and TFs often accumulate cancer driver mutations, such as TP53 (Rivlin et al. 2011) and KRAS (M. T. Wang et al. 2015), and are the targets of anti-cancer drugs (Bhagwat and Vakoc 2015; Bhullar et al. 2018).

Due to technical limitations, the study of protein signalling activities has been for many years limited primarily to the study of a few key signalling proteins at a time using antibodies, which was recently expanded to a few hundred via the use of reverse-phase protein arrays (RPPA) (J. Li et al. 2013). The Clinical Proteomic Tumour Analysis Consortium (CPTAC) has revolutionized the study of cancer proteomes, including proteins and respective post-translational modifications (PTMs), through the application of Mass Spectrometry (MS)-based proteomics (B. Zhang et al. 2019). MS-based proteomic profiling of human cancers has the potential to uncover molecular insights that might be otherwise missed by genomics- and transcriptomics-driven cancer research. CPTAC enabled to (i) identify additional cancer molecular subtypes (Mun et al. 2019; Gao et al. 2019), (ii) find that changes at the genomic and transcriptomic level are often buffered at the proteomic level (Mertins et al. 2016; B. Zhang et al. 2014; Gonçalves et al. 2017; Sousa et al. 2019) and (iii) uncover dysregulated signalling pathways by phosphoproteomics data integration (Clark et al. 2019).

In efforts to find novel therapeutic opportunities from kinase and TF oncogenic signalling, it is crucial to understand how the activities of these key signalling proteins are changing across tumours. Previous studies found that TF mutations were correlated with transcriptional dysregulation in cancer cell lines and primary tumours, and that TF activities can act as predictors of sensitivity to anti-cancer drugs (Garcia-Alonso et al. 2018). Similar results were found regarding the impact of oncogenic mutations on kinase signalling. However, these studies were focused on few kinases and cancer types (Guo et al. 2008; Guha et al. 2008; Creixell et al. 2015; Lundby et al. 2019). Despite all of these efforts, a systematic Pan-Cancer analysis of the regulation of kinase and TF activities across tumours is still lacking.

In this study, we mined multi-omics datasets from patient tumours and cancer cell lines to study the regulation of kinases and TFs across tumour types. We estimated the activities of TFs and kinases from the gene expression levels and phosphorylation changes of their targets, deriving activity profiles of 292 TFs and 218 kinases across 1,110 primary tumors from TCGA and CPTAC and 77 cancer cell lines. We used these kinase and TF activities to study the principles of regulation of these signalling proteins by mutations, changes in abundance or phosphorylation. We show how their patterns of activity co-regulation reflect underlying signalling relationships and we identify the signalling molecules that show high degree of regulation in each tumour type. Finally, we show how these TF/kinase activities can be predictive of differential survival across patients. The protein activities profiles across over 1000 patient samples serve as a resource to study the dysregulation of signalling across different tumour types.

4.3. Results

4.3.1. Standardized multi-omics pan-cancer dataset

To study the regulation of protein activities of cancer cells, we compiled and standardized multi-omics datasets made available by the CPTAC consortium (**Figure 4.1A; Methods**). These datasets comprised of cancer patient samples with matched somatic mutations, gene copy number variation (CNV), mRNA expression, protein abundance, phosphorylation and clinical data from 9 tissues: breast (Mertins et al. 2016; Koboldt, Fulton, et al. 2012), brain (Petralia et al. 2020), colorectal (B. Zhang et al. 2014; Muzny et al. 2012; Vasaikar et al. 2019), ovarian (H. Zhang et al. 2016; Bell et al. 2011), liver (Gao et al. 2019), kidney (Clark et al. 2019), uterus (Dou et al. 2020), lung (Gillette et al. 2020) and stomach (Mun et al. 2019). In addition, we collected data for breast (Lapek et al. 2017; Lawrence et al. 2015) and colorectal (Roumeliotis et al. 2017) cancer cell lines, for which multi-omics data were available (**Figure 4.1A; Methods**). In summary, we assembled a multi-omics atlas that provides the opportunity to build an integrated picture of the cancer genome, transcriptome and (phospho)proteome, with 1008 samples (932 tumours and 76 cell lines) matching all data types available per dataset.

We first calculated correlations between each protein and phosphosites that mapped to the same protein, across up to 1008 samples. Across all pairs there is a

correlation of 0.49 (*P-value* < *2.2x10e-16*), with an average protein-phosphosite correlation of 0.39 (**Supplementary figure 4.1A, Supplementary figure 4.1B**), in agreement with previous studies (Arshad et al. 2019). This result shows that phosphorylation levels are, to some extent, confounded by the corresponding protein abundance (R. Wu et al. 2011). To be able to focus on phosphorylation changes that are not driven primarily by protein abundance differences, we regressed-out matched protein abundance from the phosphorylation data in our compiled dataset (**Supplementary figure 4.1A, Supplementary figure 4.1B; Methods**).

4.3.2. Landscape of protein activities in cancer

The genomics characterization of tumour samples has so far been primarily focused on stratifying samples by their mutational profiles or changes in abundance of specific biomolecules such as transcripts, protein or phosphorylation states (Koboldt, Fulton, et al. 2012; Muzny et al. 2012; H. Zhang et al. 2016; Vasaikar et al. 2019; Clark et al. 2019). We and others have shown that changes in phosphorylation and gene expression levels can be used to infer the activation states of protein kinases and TFs (Ochoa et al. 2016; Garcia-Alonso et al. 2018). Based on these methods we set out to define the landscape of kinase/TF activity patterns across these tumour samples.

The kinase activities were estimated from the protein abundance-corrected phosphorylation data using a z-test (Hernandez-Armenta et al. 2017) (**Figure 4.1B; Methods**). Briefly, the activity of a given kinase in a sample is estimated by comparing the changes in phosphorylation of its substrates with changes of all other phosphosites. Similarly, the activation state of TFs was inferred from the changes in gene expression of their known transcriptional targets using the DoRothEA regulons (Garcia-Alonso et al. 2019) coupled with the VIPER algorithm (Alvarez et al. 2016) (**Figure 4.1B; Methods**). In total, we estimated the activities of 292 TFs across 1,187 cancer samples (1,110 primary tumours and 77 cell lines) (**Table S4.1**). For the estimation of kinase activities, we evaluated different lists of kinase substrates from repositories, computational text mining (Bachman, Gyori, and Sorger 2019), kinase inhibitor experiments (Hijazi et al. 2020) or phosphorylation of cell extracts (Sugiyama, Imamura, and Ishihama 2019) (**Supplementary figure 4.2A, Supplementary figure 4.2B**). We tested each list in a compilation of phosphoproteomic experiments where kinase regulation is known (Ochoa et al. 2016; Hernandez-Armenta et al. 2017) (**Methods**), keeping those from repositories and text-mining as the most accurate (**Supplementary figure 4.3A, Supplementary figure 4.3B**). After applying this approach,

we inferred the activities of 218 kinases across 980 samples (930 tumours and 50 cell lines) (**Table S4.1; Methods**).

For some kinases, there are phosphosites within the kinase itself that are known to activate or inhibit it. As a validation, we correlated the estimated activity scores with the quantifications of activating phosphosites, finding the expected higher correlation when compared with phosphosites without annotation (**Figure 4.1C**). A similar trend was observed when excluding the kinase auto-regulatory phosphosites before re-estimating the activities (**Supplementary figure 4.3C**). Finally, we benchmarked the kinase activity scores using reverse phase protein array (RPPA) data from the TCGA program. We first evaluated the agreement between the MS-based and the RPPA-based phosphosite quantifications. We found that phosphosite pairs corresponding to the same phosphosite show higher correlations than random pairs (**Figure 4.1D**). Then, we found that the RPPA phosphosites correlate significantly better with the activity of kinase bearing the phosphosites than with other kinase activities (**Figure 4.1E**).

The activity profiles of kinase and TFs across a large number of samples allows us to ask how these activities are themselves regulated. We first selected 99 kinases and 120 TFs that are strongly regulated in at least 5% of all samples (**Supplementary figure 4.3D**). We then correlated these activities with changes in gene copy number (CNV), mRNA and protein levels or changes in phosphorylation levels of the respective protein (**Table S4.2**). We observed that 55% of kinase activities correlated with their phosphorylation state and only 27% correlated with changes in protein abundance (**Figure 4.1F**). Contrary to this, TF activities are most often correlated with changes in abundance of the TF, as measured by RNA (91%) or protein (59%), with fewer cases of significant correlations with phosphorylation levels (29%) (**Figure 4.1F**). TF phosphosites predicted to be important for function (Ochoa et al. 2020) are more likely to show significant correlations with the TF activity (**Supplementary figure 4.3E**).

Overall, these results showed that our kinase activity estimates are likely to capture kinase regulatory events across different tumour types, and therefore the usefulness of our multi-omics atlas to study kinase signalling in cancer.

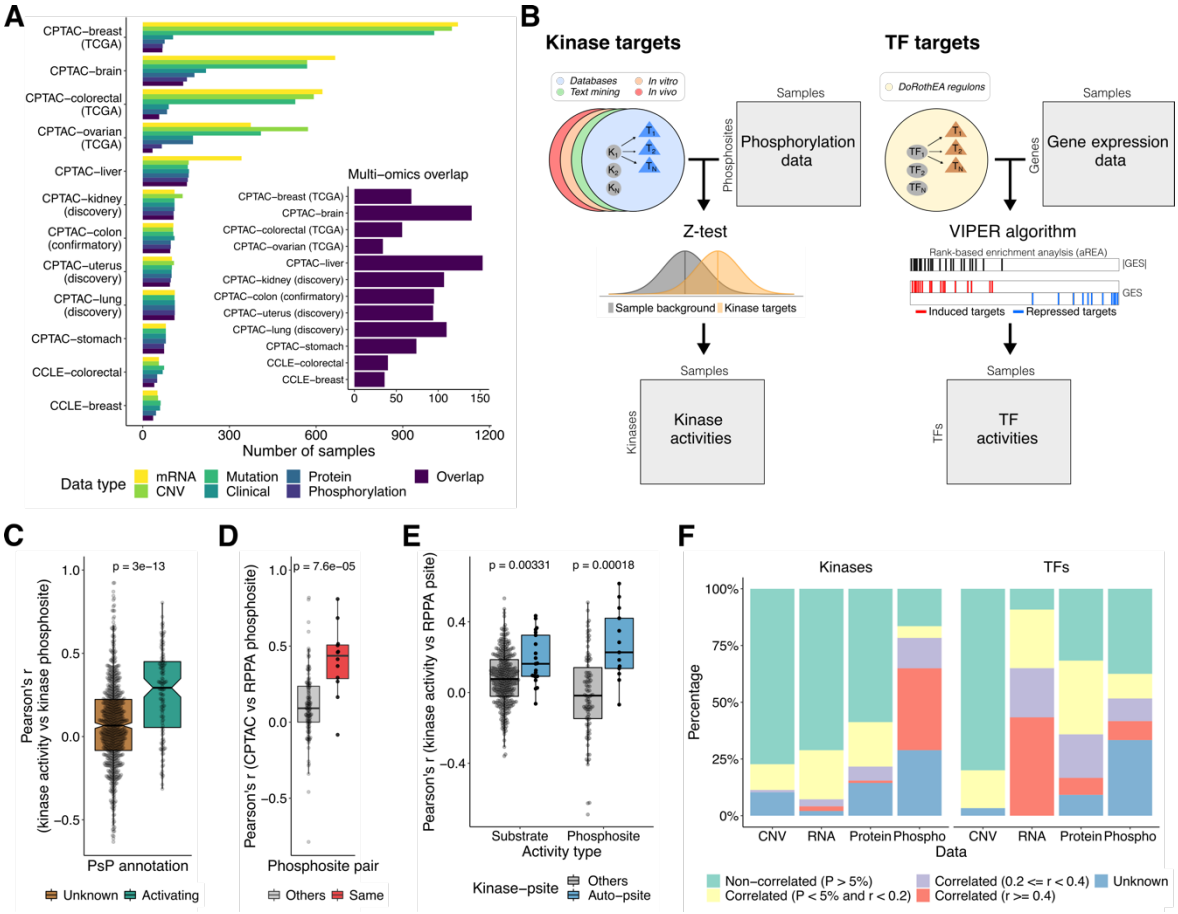


Figure 4.1. Multi-omics atlas and inference of protein activities. (A) Number of samples by cancer dataset and data type. **(B)** Schematic representation of kinase and TF activity inference. GES, gene expression signature. **(C)** Comparison of the Pearson's correlation distributions between the kinase activities and the quantifications of phosphosites (log2 fold-changes) that mapped to the same kinase, with (n = 126) and without (n = 793) annotation (activating) in PhosphoSitePlus. A P-value from a Wilcoxon rank sum test is shown. **(D)** Pearson's correlation between the CPTAC MS-based and the TCGA RPPA-based phosphosite quantifications, for the same phosphosite pair (n = 12) and others (n = 132). A P-value from a Wilcoxon rank sum test is shown. **(E)** Comparison of the Pearson's correlation between the RPPA phosphosites and the kinase activities, for kinase-phosphosite pairs mapping to the same kinase (auto-phosphosite) and other pairs. The activities were calculated using the kinase substrates (n = 336 and n = 21) and the kinase regulatory phosphosites (n = 94 and n = 13) (**Methods**). The P-values from Wilcoxon rank sum tests are shown. **(F)** Percentage of kinases and TFs significantly and not significantly correlated with the corresponding CNV, RNA, protein and phosphorylation levels. The proteins without correlations due to lack of data or reduced number of samples (n < 10) were labeled as unknown (blue).

4.3.3. Impact of genetic variation on protein abundance and activities

The large number of cancer samples in this study constitutes a resource to measure the effects of genetic alterations, i.e., somatic mutations and CNVs, on protein abundances and activities. We first set out to assess the effects of CNVs on the mRNA and protein abundances. Similarly to our previous reports, the CNVs showed a stronger correlation with the mRNA than with the protein levels (**Supplementary figure 4.4A, Supplementary figure 4.4B**), highlighting mechanisms of post-transcriptional control and gene dosage buffering at the protein level (Sousa et al. 2019; Gonçalves et al. 2017). We then extended the analysis to globally assess the effects of mutations (**Methods**), and we found that proteins carrying loss-of-function (LoF) alterations, including frameshift, nonsense, splice site and stop codon loss, caused on average a significant decrease in protein abundance. This was not observed with in-frame and missense mutations (**Supplementary figure 4.4C**). To validate the decrease of protein abundance for LoF mutations, we confirmed that this was also recapitulated in a proteomic dataset with 125 cancer cell lines (CCLs) from the NCI60 and CRC65 panels (Frejno et al. 2020) (**Supplementary figure 4.4D; Methods**). These observations confirm that the genetic alterations are often recapitulated at the protein level as captured by the MS data.

We next looked at the impact of genetic alterations on TF and kinase activity estimates. On average, we did not observe reduced activity for proteins carrying different types of mutations in the tumour samples (**Supplementary figure 4.5A**), with only a very modest average decrease in activities for frameshift mutations found in the cell line data (**Supplementary figure 4.5B**). This observation did not depend on the degree of predicted deleterious impact of the mutations (**Supplementary figure 4.5C, Supplementary figure 4.5D**) nor on the purity of the tumour samples (**Supplementary figure 4.5E**). To further characterize this unexpected result, we focused on highly mutated cancer genes. As an example, we investigated the impact of the BRAF V600E mutation on the activities of proteins from the MAPK/ERK signaling transduction pathway (**Methods**). Surprisingly, across all samples BRAF V600E mutations were not significantly associated with changes in activity of key pathway components, including BRAF itself, MAPK1, MAPK3, MAP2K1 and MAP2K2 (**Supplementary figure 4.5F**). Instead, we found that CDK1 and CDK7 were more active in samples carrying the mutation (*FDR < 5%*) (**Supplementary figure 4.5G**). This suggests that samples carrying a BRAF V600E mutation will often have kinase activity levels that have adapted to the mutational state, likely having increased proliferation, as indicated by the CDK1 levels but not a higher level of activation of the pathway.

We then extended the analysis by systematically associating the activity of kinases and TFs with the recurrent mutational status of any given gene mutated in at least 5 tumour samples (**Methods**). As seen for the BRAF example, we didn't observe any case where recurrent mutation of the kinase itself was associated with a significant change in its activity

as measured by the phosphorylation of its substrates. This indicates that there is significant adaptation of the signalling state of the cell after mutations. On the other hand, we found 193 significant associations (*FDR < 5%*) between mutations in other genes and changes in kinase activity levels (**Figure 4.2A; Table S4.3**). For example, samples with mutations on STK11 (serine/threonine kinase 11) don't show a pronounced change in activity of STK11 substrates but have decreased activity for PRKACA kinase, a known activator of STK11 (*FDR = 9.6e-5*; combined string network weight = 0.94) (**Figure 4.2C**). Other examples include increased activity for CDK1 and MAPK13 in samples with mutations in TP53, and for AKT3 when PTEN is mutated (*FDR < 5%*) (**Figure 4.2C**). Unlike for kinases, we found several cases where the mutation of a TF was associated with a change in its own activity as is the case for mutations in TP53, GATA6, SREBF2 and EBF1 (*FDR < 5%*) (**Figure 4.2B - inner plot, Supplementary figure 4.6; Table S4.3**). In addition, we found 11,128 significant associations between a mutated gene and a changed TF activity (*FDR < 5%*) (1,087 for *FDR < 1%*) (**Figure 4.2B - outer plot; Table S4.3**), including increased activity for E2F4 and TFDP1 coupled with TP53 mutation (**Figure 4.2D**).

The associations between mutated genes and altered protein activities contain several examples of previously known functional relationships. To evaluate this more broadly, we confirmed that our predicted associations were enriched in protein-protein functional associations annotated in the STRING database, both for the kinases and the TFs (*P-value < 5%*) (**Figure 4.2E, Figure 4.2F - top plots**). We also performed an enrichment analysis using the string network along multiple cutoffs of adjusted P-values (**Methods**). The -log10 transformed P-values from the enrichment test increased as the association cutoffs were incremented (**Figure 4.2E, Figure 4.2F - bottom plots)**, validating the generality of the significant associations. Overall, the genetic associations found are enriched in previously known functional associations, containing potential novel regulatory relationships for future experimental exploration.

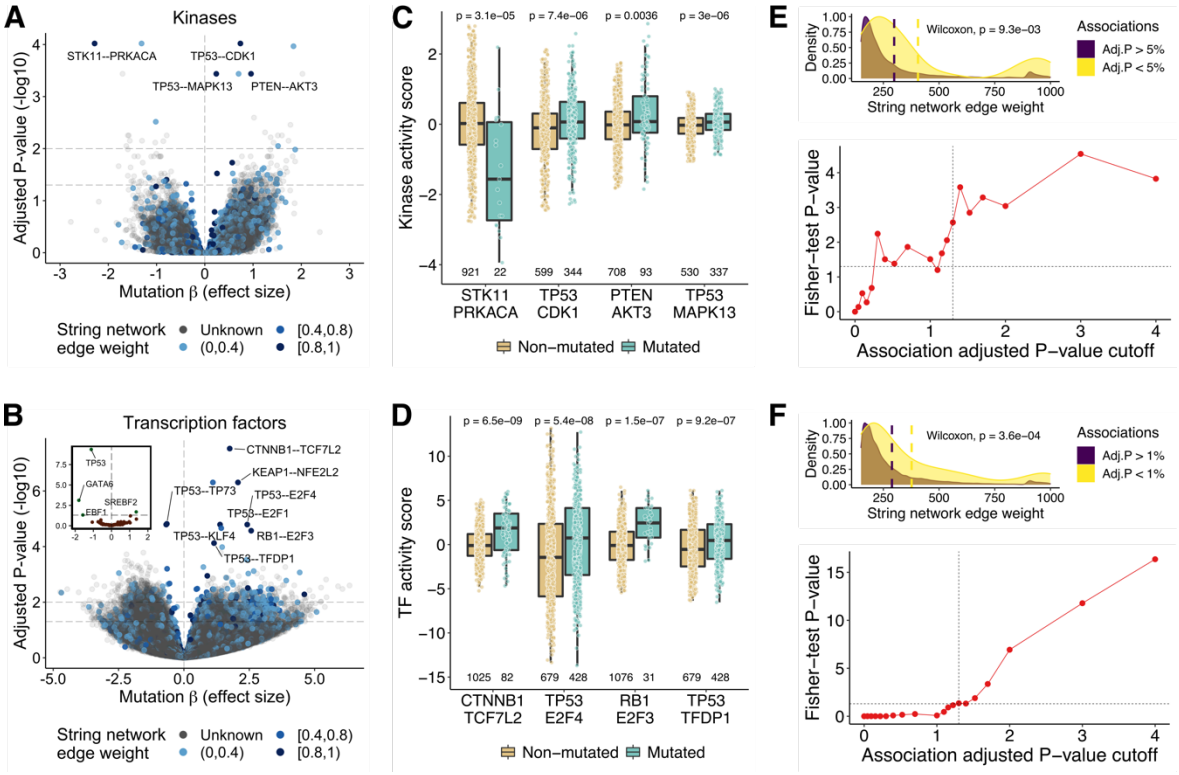


Figure 4.2. Genetic associations. (A) Volcano plot displaying the associations between the mutational status of genes and the activity of kinases. The x-axis contains the mutation coefficient (effect size) and the y-axis the adjusted P-values. The associations are represented in the form of a mutated gene - kinase. The color gradient represents the string network edge weight interval of the pair (grey if the pair is not in the string network). **(B)** Same as (A) for the TFs. The inner plot shows the effects of TF mutations on their own activities. **(C) (D)** Examples of the genetic associations highlighted in the volcano plots. The x-axis represents the associations and the y-axis the protein activities. The colors stratify the samples by their mutational status in the respective genes. The outliers (defined as the data points beyond Q1-1.5*IQR and Q3+1.5*IQR, where Q1 and Q3 are the first and third quartiles and IQR is the interquartile range) were removed from the distributions for representation purposes. The number of protein activity quantifications (including outliers) are shown beneath each boxplot. The P-values from Wilcoxon rank sum tests comparing both distributions are shown. All data points (including outliers) were used to calculate the P-values. **(E) Top panel.** Density plots comparing the edge weight distributions in the string network of the significant and non-significant association pairs obtained with the kinases. **(E) Bottom panel.** Enrichment of the associations in the string network (edge weight > 850) along multiple cutoffs of statistical significance. The x-axis shows the adjusted P-value cutoffs (-log10) and the y-axis the Fisher-test P-values (-log10). **(F)** Same as (E) for the TFs.

4.3.4. An atlas of kinase and TF regulation in cancer

The estimation of kinase and TF activities across a large set of tumour samples from different tissues provides a first look at the space of tumour signalling states as measured by hundreds of regulators. We projected the activity profiles in a lower-dimensional space

using the uniform manifold approximation and projection for dimension reduction algorithm (UMAP) (**Methods**). For both the kinases and TF activities we observed that cancer samples were not clustered by experimental study (**Figure 4.3A, Supplementary figure 4.7A**). The same was also observed using a principal component analysis (PCA) (**Supplementary figure 4.7B, Supplementary figure 4.7C**). These results suggest that our normalization procedures helped to mitigate the technical biases between studies, being likely superimposed by biological variation.

After including only one kinase from sets of redundant kinases based on shared substrates such as AKT1/AKT2 (**Methods**), we selected the 30 kinases with the largest amount of variation along the samples (SD > median SD). As expected, these kinases are highly correlated with the UMAP projections (**Figure 4.3B**). This set of kinases contains known cancer drivers and kinases with inhibitors already used in the clinic as cancer treatment, such as BRAF, AKT, MAP2K1, SRC among others. Examining the tumour samples in this two-dimensional representation indicates that highly regulated kinases in the same pathway tend to be activated or inhibited across the same samples (**Figure 4.3C**). For example, we found that tightly co-regulated kinases from the MAPK signalling pathway, including PRKCB (PKC), BRAF (RafB), MAP2K1 (MEK1) and MAPK1 (ERK), share the same activation pattern along the cancer samples (**Figure 4.3C**). CDK1 is known to phosphorylate the casein kinase 2 (CSNK2A1) (Bosc et al. 1995). These kinases together showed opposite correlations with the UMAP projections and, consequently, a distinct regulatory state across the samples (**Figure 4.3B, Figure 4.3C)**. To study this more formally, we obtained pairwise kinase regulatory relationships deposited in the OmniPath database (Türei, Korcsmáros, and Saez-Rodriguez 2016) and correlated their activities (**Methods**). We found that kinases that regulate each other were more likely to have correlated patterns of activity across samples (**Figure 4.3D**). This was still observed when taking into account cases where the pair of kinases shared some substrates (**Supplementary figure 4.7D; Methods**). Similarly, we would expect that kinases and TFs within the same pathway will tend to have similar patterns of activation across the samples. To investigate this, we modelled the TF activities as a function of the kinase activities using linear regressions (**Methods**), identifying 5,712 significant associations at an *FDR < 5%* (3,130 for *FDR < 1%*) (**Figure 4.3E; Table S4.4**). These associations were enriched in known kinase-TF functional interactions (**Supplementary figure 4.7E, Supplementary figure 4.7F**), including for example the relation between CDK1 activity and the activities of E2F4 and TFDP1 (Spring et al. 2017; Jiao et al. 2017) (**Supplementary figure 4.7G**). Altogether, these results corroborate that the variation in activities across the samples is shaped to some extent by the underlying regulatory relationships.

Our analysis can indicate the kinases that are most often dysregulated in cancer. For comparison, we also estimated kinase activity changes from phosphoproteomic measurements in a large panel of other conditions (Ochoa et al. 2016). We observed a correlation between the degree of regulation of kinases in cancer and non-cancer conditions (*r = 0.78, P-value < 2.2e-16*), with AKT1 and the cell-cycle kinases CDK1/2 and AURKB being highly regulated in both sets of conditions (**Figure 4.3F**). Kinases deviating from the regression line can be classified as preferentially regulated in the tumours or in the other conditions (**Methods**). There were a larger number of kinases specifically dysregulated in cancer (e.g., including PRKACA, CSNK2A1 and MAPK1) compared with other non-cancer conditions (**Figure 4.3F**). The kinases MAPKAPK2, RPS6KB1 and RPS6KA3 were less likely to be dysregulated in the tumours when compared to degree of regulation in other perturbations (**Figure 4.3F**). We performed the same analysis by tissue type (**Supplementary figure 4.8A**). The number of specifically dysregulated kinases was consistently higher in the tumours than the non-cancer conditions in all tissues (**Supplementary figure 4.8A, Supplementary figure 4.8B**). The inter-tissue variation regarding the number of dysregulated kinases in tumours correlated with the number of samples but not with the number of kinases quantified in the tissues (**Supplementary figure 4.8C**). Some kinases (e.g., PRKACA, CSNK2A1 and MAPK1) were found specifically dysregulated across multiple tumour types, but more than half were dysregulated in just one tissue (68%) such as MYLK kinase in stomach cancer and PIM3 in kidney cancer (**Supplementary figure 4.8D**).

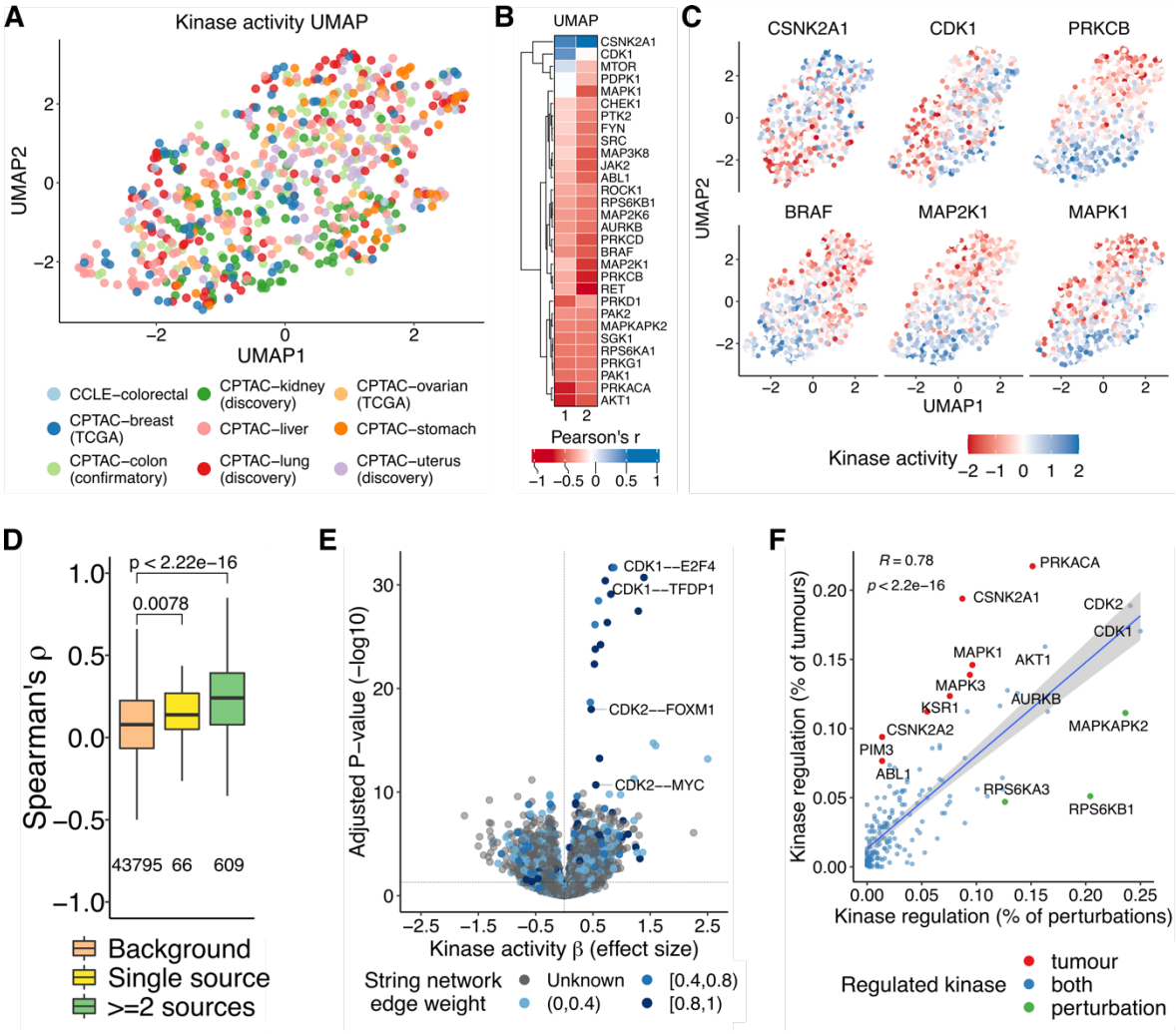


Figure 4.3. Regulation of protein activities in tumours and human perturbations. (A) UMAP projection of the kinase activity matrix (kinases as variables). The samples are coloured by experimental study. **(B)** Pearson correlation coefficient between the UMAP projections and the activity of non-redundant highly variable kinases. **(C)** Kinase activity gradient along the samples for a selection of the kinases shown in (B). **(D)** Spearman's rank correlation coefficients between the activities of kinases known to co-regulate each other. The pairwise kinase co-regulatory relationships were obtained from the OmniPath database and stratified by their presence in the OmniPath's sources (as single source or in at least two different sources). We only kept activating and consensual interactions along the sources. The background corresponds to kinase pairs without known co-regulation events. The distributions were compared to the background using Wilcoxon rank sum tests. **(E)** Associations between the activity of kinases and TFs. The x-axis contains the kinase coefficients (effect sizes) and the y-axis the adjusted P-values. Each association is represented in the form of kinase - TF. The colour gradient represents the edge weight of the pair in the string network (grey if not present). **(F)** Linear regression between the percentage of samples where the kinase is regulated in the perturbed conditions and in the tumour samples. The trend line and the Pearson's r, with the respective P-value, are shown. In red and green are the kinases preferably regulated in the tumours and in the conditions, respectively. In blue are the kinases regulated in both.

4.3.5. Differential protein activity is associated with changes in patients survival

Survival analyses from multi-omics datasets have been largely based on gene or protein expression differences between groups of patients (Ally et al. 2017; Bell et al. 2011; Petralia et al. 2020; Gao et al. 2019). However, kinase and TF activities should better capture the signalling state of the cancer samples and could be also linked to overall patient survival (OS). To explore this, we first performed a log-rank test to compare the Kaplan-Meier (KM) survival curves between patients with TF and kinase activities classified as inactive, neutral and active (**Methods**). We found several TFs and kinases significantly associated with OS in different tumour types (**Figure 4.4A, Figure 4.4B; Table S4.5**). For instance, the degree of MYC activity was correlated with OS in brain and liver cancers (**Figure 4.4C, Figure 4.4D**). In both cases, patients with high MYC activity showed less OS than patients with neutral and inactivated MYC (**Figure 4.4C, Figure 4.4D**). According to the literature, MYC overexpression is a poor prognosis factor in liver and paediatric brain tumours (Lin et al. 2010; K. Zheng, Cubero, and Nevzorova 2017; Hutter et al. 2017). To take into account the effects of possible confounding covariates, we performed a multivariate Cox regression analysis using the protein activity scores as a predictor, while controlling for conventional clinical covariates and the genotype of recurrently mutated genes (**Methods**). Reassuringly, our findings with the log-rank tests were largely recapitulated with the Cox models (*brain: hazard ratio (HR) = 1.50 (95% CI 1.27-1.76), adjusted P-value = 2.6e-5; liver HR = 1.17 (95% CI 1.06-1.31), adjusted P-value = 1e-2*). Interestingly, we also found that high activity of FOXA1 and FOXM1 is a good and poor prognostic factor in liver cancer, respectively (*FOXA1: HR = 0.69 (95% CI 0.55-0.85), adjusted P-value = 4.7e-3; FOXM1: HR = 1.39 (95% CI 1.17-1.66), adjusted P-value = 1.9e-3*) (**Supplementary figure 4.9A, Supplementary figure 4.9B**). These two proteins are known for their opposite role in hepatocarcinogenesis. On the one hand, elevated expression of FOXM1 promotes tumour cell proliferation and, on the other hand, FOXA1 inhibits tumour progression by suppression of PIK3R1 expression (He et al. 2017; M. Yu et al. 2016).

Regarding the kinases, we found that elevated activity of MAP3K8 and PRKCA was associated with less probability of survival in renal cancer (*MAP3K8: HR = 10.5 (95% CI 5.81-19), adjusted P-value = 1.1e-13; PRKCA: HR = 6.15 (95% CI 4.27-8.84), adjusted P-value = 2.3e-21*) (**Figure 4.4E, Figure 4.4F**). In agreement with these results, the overexpression of both protein kinase C and mitogen-activated protein kinase 8 has been associated with a higher invasiveness of kidney tumours (Engers et al. 2000; J. Li and Gobe 2006; F. Liu et al. 2016; Su et al. 2015). Lastly, overactive AURKB was correlated with a

lower survival rate in breast cancer (*HR = 27.2 (95% CI 11.2-66.2), adjusted P-value = 5.1e-12*) (**Supplementary figure 4.9C**), as previously found at the gene expression level (D. Huang et al. 2019).

Altogether, these results indicate that the inference of kinase and TF activities can be a relevant prognostic tool in cancer studies.

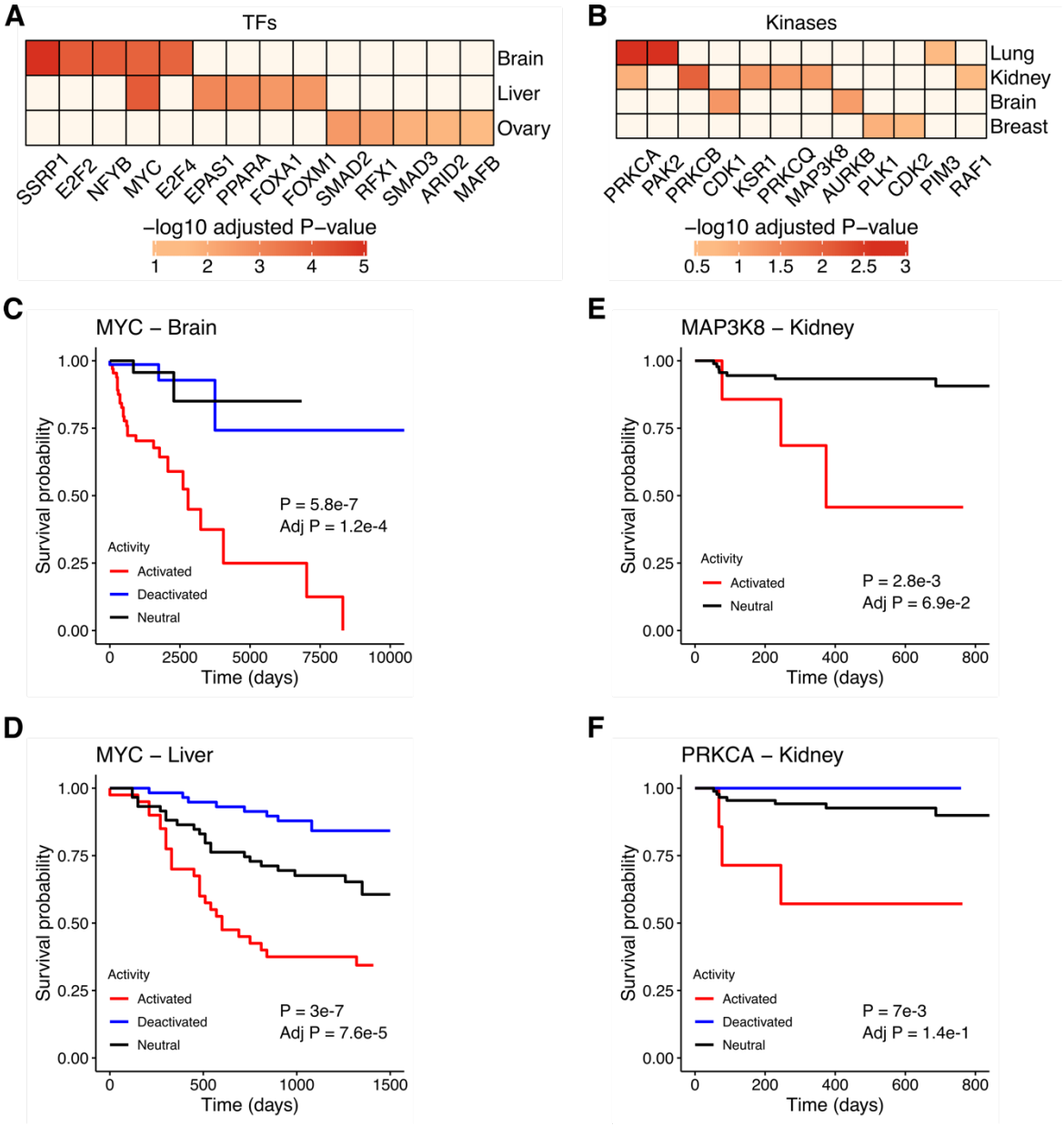


Figure 4.4. Survival analysis using kinase and TF activities. (A) Heatmap of log-rank test adjusted P-values (-log10) comparing Kaplan-Meier survival curves between cancer samples with TF activities classified as inactive, neutral and active (top 5 associations per tissue; *FDR < 5%*). **(B)** Same as (A) for the kinases (associations with *FDR < 20%*). KM survival plots comparing the survival probabilities (y-axes) as a function of time in days (x-axes) for MYC in **(C)** brain (inactive = 71, neutral = 39, active = 67) and **(D)** liver (58, 59, 40) cancer and **(E)** MAP3K8 (neutral = 95, active = 7) and **(F)** PRKCA (inactive = 3, neutral = 92, active = 7) in renal cancer. The log-rank P-values are shown in the plots.

4.4. Discussion

Kinases and TFs are important mediators of cell signalling regulation and sensitivity to anti-cancer drugs. Here, we have compiled multi-omics datasets made available by the TCGA and CPTAC consortia and cell line studies and we were able to estimate the activities of 218 kinases and 292 TFs across 1,110 primary tumours and 77 cancer cell lines. Based on these we found that kinase activities appear to be primarily regulated by phosphorylation level with fewer cases of significant correlation with the predicted kinase protein abundance levels. Contrary to this, the predicted TF activity is primarily correlated with the mRNA/protein level of the TF itself with a smaller proportion of TFs with significant correlations with the phosphorylation state. This difference in regulation is not simply due to lack of detection of phosphosites, as TFs have a median value of 6 phosphosites detected compared to 9 for kinases. A larger fraction of the TF activities is correlated with their mRNA levels than the protein abundance. This result is non-intuitive since the protein abundance should be a better proxy for activity. It is possible that this is due to the fact that TF activities are derived directly from the same mRNA datasets while there will be some degree of technical variation due to sample preparation and analysis when compared with the protein dataset.

Intuitively, the activity of a given protein (i.e., kinases and TFs) might be positively or negatively affected by mutations in the same protein or in other proteins it interacts with throughout the signalling networks. Our genetic analysis identified associations between mutated genes and the activities of kinases and TFs which were significantly enriched for known protein-protein interactions. Moreover, we found that the activities of the transcription factors TP53 and SREBF2 were correlated with their mutational status, as previously described (Garcia-Alonso et al. 2018). Nevertheless, we did not observe a general correlation between deleterious mutations within a kinase/TF and its activity including expected associations such as the BRAF V600E mutation and the activities of BRAF itself or other members of the MAPK pathway. These results were not due to issues linked with purity or immune infiltration as the same was observed in cell lines. Nevertheless, there may be issues in linking mutations with signalling differences due to single-cell-level variances that can not be systematically profiled in bulk (Lun and Bodenmiller 2020). In the future, single-cell multi-omics profiling may allow us to consider intra-tumour genetic heterogeneities. We speculate that these observations are more likely explained by

feedback loop mechanisms that are prevalent in signalling pathways (Lito, Rosen, and Solit 2013). These results emphasize the difficulty in interpreting the impact of mutations on signalling networks and the importance of studying directly the dysregulation of signalling in cancer (Yaffe 2019).

Known kinase regulatory pairs have strong patterns of co-regulation across the compiled dataset. These results suggest that the kinase activity estimates are meaningful, and that the variation in kinase activities along cancer samples is likely driven by biological factors. We showed in a previous work that these co-regulation signals can be used to predict kinase regulatory networks (Invergo et al. 2020). Similarly, we found many significant associations between the activities of kinases and TFs, which were significantly enriched in known functional interactions. This indicates that this compendium of protein activities may be useful in the future development of methods to reconstruct the signalling networks. Nevertheless, even the strongest correlations were modest in aggregate: well-studied kinase-kinase regulatory pairs showed a median correlation of their predicted activities of 0.25 (**Figure 4.3D**). Our prior knowledge about kinase co-regulation is currently limited at multiple levels, including yet to be found kinase regulatory relationships and the extent that these regulatory relationships depend on the tissue of origin or other factors. We speculate that these and other confounding factors could explain the weak kinase-kinase correlations.

By comparing the kinases most often differentially regulated across tumours and after more acute perturbations, we have shown that most often regulated kinases are the same in both contexts. These include kinases such as CDK1, AKT1 and AURKB. The most plausible explanation for this would be that kinases that are often regulated in acute perturbations are directly linked to the regulation of growth and cell-cycle and other critical processes needed to be regulated in cancer cells. Alternatively, we have shown that these highly regulated kinases occur in very central positions in the signalling network (Ochoa et al. 2016) and that it is possible that some degree of regulation of these kinases is almost unavoidable. Interestingly, this comparison allowed us to identify kinases which show higher differential regulation in tumours than acute perturbation such as PRKACA, MAPK1 and MAPK3.

Finally, we show how the estimated protein activity can be linked to differences in patient survival. Given that the activities of kinases and TFs can often be estimated via antibodies targeting regulatory phosphosites, it may be possible to develop biomarkers based on these findings. In addition, kinases are very trackable drug targets with multiple kinase drugs already used to treat cancer patients. While further studies in cell based and animal models will be required to evaluate the significance of the findings presented here,

this work provides kinase activities linked to specific tumour types and mutational contexts that could be pursued for potential treatment.

4.5. Methods

Data collection

Proteomics and phosphoproteomics

The mass spectrometry (MS)-based protein and phosphosite quantifications (absolute [phospho]peptide intensities and ratios relative to controls) for the cancer samples of brain (Petralia et al. 2020), breast (Mertins et al. 2016), colorectal (B. Zhang et al. 2014), kidney (Clark et al. 2019), liver (Gao et al. 2019), lung (Gillette et al. 2020), ovarian (H. Zhang et al. 2016), stomach (Mun et al. 2019) and uterus (Dou et al. 2020) were downloaded from the CPTAC data portal (proteomics.cancer.gov/data-portal). For the colon cancer samples (Vasaikar et al. 2019), we downloaded the data from the linkedomics database (linkedomics.org/login.php). The same data for the cancer cell lines of breast and colorectal tumours was downloaded from the respective publications (Lapek et al. 2017; Lawrence et al. 2015; Roumeliotis et al. 2017). The proteins and phosphosites were identified using gene symbols, in a process described by the common data analysis pipeline (CDAP) from CPTAC. Additionally, we downloaded normalized RPPA protein and phosphorylation quantification data (183 features across 7,694 samples from 31 TCGA tumours) from the TCPA (J. Li et al. 2013) database.

Transcriptomics

The RNA-seq data was obtained in the format of read counts and Fragments Per Kilobase of transcript per Million mapped reads (FPKM). The data for the tumour tissues of breast (Mertins et al. 2016), colorectal (B. Zhang et al. 2014), kidney (Clark et al. 2019), lung (Gillette et al. 2020), ovarian (H. Zhang et al. 2016) and uterus (Dou et al. 2020) was downloaded from the GDC portal (portal.gdc.cancer.gov/). The data for the brain cancer was compiled from the paediatric cBioPortal (pedcbioportal.kidsfirstdrc.org/); for the liver (Gao et al. 2019) from NODE (www.biosino.org/node/) (accession ID: OEP000321); for the stomach (Mun et al. 2019) from GEO (ncbi.nlm.nih.gov/geo/) (accession ID: GSE122401);

and for the colon cancer (Vasaikar et al. 2019) from the authors. The cancer cell lines (Lapek et al. 2017; Lawrence et al. 2015; Roumeliotis et al. 2017) data was downloaded from the CCLE data portal (portals.broadinstitute.org/ccle/data).

Genomics - somatic mutations

The whole genome sequencing (WGS)-derived somatic mutations for the brain cancer samples (Petralia et al. 2020) were downloaded from the paediatric cBioPortal (pedcbioportal.kidsfirstdrc.org/) in Mutation Annotation Format (MAF) files. For the breast (Mertins et al. 2016), colorectal (B. Zhang et al. 2014) and ovarian (H. Zhang et al. 2016) cancers, the whole exome sequencing (WES)-derived MAF files were downloaded from the cBioPortal (cbioportal.org). The MAF file for the colon cancer samples (Vasaikar et al. 2019) was downloaded from the linkedomics database (linkedomics.org/login.php). Regarding the kidney (Clark et al. 2019), lung (Gillette et al. 2020) and uterus (Dou et al. 2020) cancers, we downloaded the MuTect2-called and VEP-annotated VCF files from the GDC data portal (portal.gdc.cancer.gov/). For the liver (Gao et al. 2019) and stomach cancers (Mun et al. 2019), we obtained the somatic mutations from the publication and authors, respectively. The mutation data for the colorectal and breast cancer cell lines (Lapek et al. 2017; Lawrence et al. 2015; Roumeliotis et al. 2017) was obtained from the DepMap portal (depmap.org/portal/).

Genomics - somatic copy number alterations

The somatic copy-number variation (CNV) data was downloaded as discretized GISTIC2 scores (Beroukhim et al. 2007; Mermel et al. 2011) and segment-level log2 ratios between the tumour and normal samples. The GISTIC2 scores can be -2 (strong copy-number loss, likely a homozygous deletion), -1 (shallow deletion, likely a heterozygous deletion), 0 (diploid), 1 (low-level gain of copy number, generally broad amplifications) and 2 (high-level increase in copy number, often focal amplifications). The GISTIC2 scores for the tumour samples of breast (Mertins et al. 2016), colorectal (B. Zhang et al. 2014) and ovarian (H. Zhang et al. 2016), and for the cancer cell lines (Lapek et al. 2017; Lawrence et al. 2015; Roumeliotis et al. 2017) were downloaded from the cBioPortal (cbioportal.org). The same data for the brain cancer samples (Petralia et al. 2020) was downloaded from the paediatric cBioPortal (pedcbioportal.kidsfirstdrc.org/); for the colon samples (Vasaikar et al. 2019) from linkedomics (linkedomics.org/login.php); and for the liver samples (Gao et al. 2019) from NODE (biosino.org/node/index) (accession ID: OEP000321). The segment-level

log2 ratios for the kidney (Clark et al. 2019) and uterus (Dou et al. 2020) cancer samples were provided by the authors of the respective publications.

Clinical data

The metadata and clinical information from the patients of breast (Mertins et al. 2016), colorectal (B. Zhang et al. 2014) and ovarian (H. Zhang et al. 2016) cancers was obtained from the CPTAC (proteomics.cancer.gov/data-portal) and the cBioPortal (cbioportal.org) databases. The survival data for these patients was collected from (J. Liu et al. 2018), whereas the cancer subtypes were obtained from the respective publications and also using the *PanCancerAtlas_subtypes* function from the *TCGAbiolinks* R package (Colaprico et al. 2016). For the brain (Petralia et al. 2020), kidney (Clark et al. 2019), liver (Gao et al. 2019), lung (Gillette et al. 2020), stomach (Mun et al. 2019) and uterus (Dou et al. 2020) cancers, we downloaded the clinical information from the CPTAC portal and from the respective publications. For the colon cancer samples, we downloaded the data from the linkedomics database (inkedomics.org/login.php). The clinical data from the cancer cell lines donors (Lapek et al. 2017; Lawrence et al. 2015; Roumeliotis et al. 2017) was obtained from the CCLE data portal (portals.broadinstitute.org/ccle/data). Altogether, we collected the following information about the cancer patients: age, gender, ethnicity, race, height, weight, cancer histological type and subtype, tumour stage, overall survival and survival time in days.

Data pre-processing and normalization

Proteomics

The label-free protein quantifications (precursor areas) for the colorectal tumours (B. Zhang et al. 2014) and the tandem mass tag (TMT) protein intensities for the breast and colorectal cancer cell lines (Lapek et al. 2017; Lawrence et al. 2015; Roumeliotis et al. 2017) were pre-processed and transformed to log2 fold-changes as previously described (Sousa et al. 2019). For the brain (Petralia et al. 2020), lung (Gillette et al. 2020) and stomach (Mun et al. 2019) cancers, the sample replicates were combined by averaging the log2 fold-change values of each protein. After that, we removed 6 outlier samples from colorectal cancer with an absolute median log2 fold-change distribution higher than 1 (2-fold). Altogether, we assembled a matrix with 14,742 proteins and 1,266 samples (1,170 cancer

samples and 96 cell lines) belonging to 9 different tissues. This matrix contained 9,941,918 protein measures (8,721,454 missing values) and 5,052 proteins quantified in at least 80% of the samples.

Phosphoproteomics

The phosphorylation measures were acquired at the phosphosite level. Each phosphosite is identified by a given protein, position and residue. The phosphosites from the different datasets were harmonized against a common reference by only keeping the phosphorylation sites that mapped correctly to the Ensembl human proteins (GRCh37 - release 98). As the phosphorylation sites were annotated at the gene symbol level (see data collection above), we mapped the phosphosites to the protein sequences using the canonical transcripts from UniProt (github.com/mskcc/vcf2maf/blob/main/data/isoform_overrides_uniprot). Duplicated phosphosites, arising from multiple phosphopeptide intensities mapping to the same phosphosite, were reduced to a single phosphosite if the log2 fold-change values were the same across all samples from the respective experimental study. All duplicated phosphosites were discarded otherwise. For the colorectal cancer cell lines (Roumeliotis et al. 2017), the relative TMT intensities (obtained by dividing the TMT intensities per the mean TMT intensity for each protein) were divided by 100 and transformed to log2. For the brain (Petralia et al. 2020), breast (Mertins et al. 2016), lung (Gillette et al. 2020) and stomach (Mun et al. 2019) cancers, the sample replicates were combined by averaging the log2 fold-change values of each phosphosite. We removed 52 outlier samples with an absolute median log2 fold-change distribution higher than 1 (2-fold). Then, the log2 fold-change distributions across samples were quantile normalized in order to ensure comparable distributions, using the *normalizeQuantiles* function from the *limma* R package (Ritchie et al. 2015). To detect phosphorylation changes that are independent of the protein abundance, we regressed-out the protein levels from the respective phosphosites using a multiple linear regression model. The phosphosite log2 fold-changes were set as the dependent variables while the protein log2 fold-changes, age and gender were set as the independent variables. The residuals from the linear model were the phosphorylation changes not driven by the protein abundance or other confounding effects (age and gender). The final phosphoproteomic matrix contained 86,044 phosphosites across 980 samples (930 cancer samples and 50 cell lines) from 9 different tissues. Due to the sparseness of the phosphorylation data (7,280,101 measures and 77,043,019 missing values), only 256 phosphosites were quantified in at least 80% of the samples (2,438 in

50%). For the downstream analyses we only considered the phosphosites (69,599) that were quantified from phosphopeptides phosphorylated at single positions.

Transcriptomics

The RNA-seq data (FPKMs and read counts) downloaded from the GDC and GEO websites (see data collection above) were converted to tabular formats using in-house R scripts. For the liver (Gao et al. 2019) and stomach (Mun et al. 2019) cancer samples, we calculated FPKM expression values from the RSEM (B. Li and Dewey 2011) expected counts using the *rpkm* function from the *edgeR* R package (Robinson, McCarthy, and Smyth 2009). We obtained the gene lengths by calculating the size (in base pairs) of the merged exons of each gene, using the *gtftools* python script (H.-D. Li 2018) (genemine.org/gtftools.php) and the GENCODE v19 human gene annotation (gencodegenes.org). After selecting the protein-coding genes, as described in the GENCODE v19 annotation, we removed the genes without expression (FPKM > 0) in at least 50% of the samples of the respective dataset. The FPKMs of each gene were subsequently log2 transformed (adding a pseudocount of 1 to avoid taking the log of 0) and converted to log2 fold-changes by subtracting the log2 median FPKM across samples. The log2 fold-changes were calculated for each dataset separately. The final gene expression matrix contained 17,056 genes across 1,187 samples and 9 tissues (1,110 cancer samples and 77 cell lines). 14,966 genes were expressed in at least 80% of the samples.

Genomics - somatic mutations

We processed the VEP-annotated VCF files from the kidney (Clark et al. 2019), lung (Gillette et al. 2020) and uterus (Dou et al. 2020) cancer samples using the *bcftools split-vep* plugin (samtools.github.io/bcftools/howtos/plugin.split-vep.html) with the following parameters: *-f '%CHROM\t%POS\t%REF\t%ALT\t%QUAL\t%FILTER\t%CSQ\n' -d -A tab*. Only the mutations passing all quality filters (FILTER == “PASS”) were selected for downstream analyses. The mutations from these samples were then collected in a single text file using in-house bash scripts. In all datasets, we selected the mutations that were annotated using the canonical UniProt transcripts (github.com/mskcc/vcf2maf/blob/main/data/isoform_overrides_uniprot) and classified as frameshift and in frame insertions/deletions (Indels), missense, nonsense, stop codon loss (readthrough mutations) and splice site. All mutations (except splice site) were standardized against the Ensembl human proteins (GRCh37 - release 98) by filtering out those mutations whose reference (wild type) residues did not match the protein sequences in the mutation

positions. The reference/mutated residues and protein position of the mutations were extracted from the HGVSp codes. In total, we collected 284,882 mutations in 17,305 protein-coding genes, across 1,168 samples (1,079 tumours and 89 cell lines) from 9 different tissues.

Genomics - somatic copy number alterations

GISTIC2 (version 2.0.23) was used to process the segment-level log2 ratios for the kidney (Clark et al. 2019) and uterus (Dou et al. 2020) cancer samples and define the gain/loss events of each gene (see data collection above), using the default parameter settings (*-genegistic* and *-savegene* parameters were both set to 1). After obtaining the discretized GISTIC2 CNV scores for each dataset, only protein-coding genes were selected as described in the GENCODE v19 human gene annotation. The final CNV matrix contained 16,520 genes across 1,025 samples and 7 tissues (947 tumours and 78 cell lines).

Normalization of gene and protein expression data

The confounding factor related to the experimental batch (e.g., CPTAC-breast, CPTAC-brain, etc.) was removed using a linear regression model. This model was implemented with the mRNA expression or protein abundance of a given gene as a dependent variable and the experimental batch as independent variable. The residuals from the linear model were the protein or mRNA variation not driven by the technical differences between cancer datasets.

NCI60 and CRC65 cell lines - data collection and pre-processing

The proteomics and phosphoproteomics data for the NCI60 and CRC65 cancer cell lines (trypsin-digested version) were downloaded from (Frejno et al. 2020). The phosphorylation residues were obtained by mapping the position of the modifications to the UniProtKB/Swiss-Prot canonical and alternative (isoforms) protein sequences (release 2020_02) (uniprot.org/downloads). Only the phosphosites mapping to serine, threonine and tyrosine residues were selected. The log10 transformed phosphorylation and protein absolute abundances (iBAQ) were set to the original values using powers of 10 (10^*abundance*). Phospho(peptide) abundances mapping to the same phosphosite or protein were averaged per sample. The absolute abundances were converted to relative

values (fold-changes) by calculating the log2 ratio of the abundances over the median abundance across cell lines. This process was performed for both cancer cell line sets. To detect net phosphorylation changes we regressed-out the protein levels from the respective phosphosites using the residuals of a linear regression model (y ~ x) where the phosphosites were set as dependent variables (y) and the proteins as independent variables (x). In total, we assembled 11,940 proteins and 45,557 phosphosites across 125 cell lines (60 from NCI60 and 65 from CRC65).

The gene expression data was downloaded in the format of FPKMs (discover.nci.nih.gov/cellminer/) and Transcripts Per Million (TPMs) (depmap.org/portal/), for the NCI60 and CRC65 cell lines, respectively. Both gene expression measures were log2 transformed and converted to fold-changes by subtracting the log2 median FPKM/TPM across cell lines. In total, we calculated the log2 fold-changes of 18,291 genes across 95 cell lines (60 and 35 from the NCI60 and CRC65 sets, respectively).

The whole genome mutation data was downloaded from the CellMiner (discover.nci.nih.gov/cellminer/) and the DepMap databases (depmap.org/portal/) for the NCI60 and CRC65 cancer cell lines, respectively. Across the NCI60 cell lines, we selected those mutations where more than 50% of the respective reads contained the alternative allele. In both datasets, we selected the mutations annotated as silent, missense, nonsense, stop codon loss, frameshift and in frame Indels. The protein position of the mutations and respective reference/mutated residues were obtained from the HGVSp codes. Altogether, we collected 585,904 mutations in 17,259 genes along 96 cell lines (60 from NCI60 and 36 from CRC65).

Inference of kinase and TF activities

Kinase and TF activities were estimated using known kinase and TF regulatory targets. The kinase-substrate relationships were obtained from (i) ProtMapper (Bachman, Gyori, and Sorger 2019), a literature-based resource of kinase substrates annotated at the phosphosite level. The resource contains phosphorylation sites aggregated from five databases (BEL Large Corpus, NCI-PID, PhosphoSitePlus, Reactome and SIGNOR) and three text-mining tools (REACH, RLIMS-P and Sparser) and (ii) a collection of phosphosites derived from *in vivo* (Hijazi et al. 2020) and *in vitro* (Sugiyama, Imamura, and Ishihama 2019) experiments. Only the phosphorylation sites correctly mapped to the Ensembl human proteins (GRCh37 - release 98) were considered for the subsequent analyses. In total, we collected the phosphorylation targets of 573 kinases.

The transcriptional targets of the TFs were compiled from the DoRothEA R package (v1.2.0), using only interactions annotated with confidence A, B and C.

The kinase activities were inferred using a one sample z-test, which was shown to perform well (Hernandez-Armenta et al. 2017). The activity of a given kinase in a given sample was estimated as follows:

Equation 4.1: $z = \frac{x - \mu}{\sigma / \sqrt{N}}$

where z corresponds to the z-score, x the average log2 fold-change of the kinase substrates, μ the average log2 fold-change of all phosphosites measured in the sample (background), $\sqrt{N}$ the square root of the number of kinase substrates (N) and σ the standard deviation of the background. Then, the z-score was used to calculate a two-tailed P-value using the *pnorm* R function [2×pnorm(-abs(z))], which was further log10 transformed and signed based on the position of the z-score in the standard normal distribution. If the z-score was in the right part of the distribution, i.e., positive, the kinase substrates showed an increase in phosphorylation in comparison to the sample background. Thus, the activity of that kinase was also expected to be increased (positive) in that sample, and vice-versa. This process was repeated for all kinases across all samples. For the downstream analyses we selected the kinase-substrate interactions from the databases and text-mining resources and the kinase activities quantified with 3 or more substrates, resulting in 218 kinases with activity estimates in an average of 437 cancer samples (of a total of 980 samples).

We also estimated kinases activities based on the phosphorylation changes of phosphosites mapping to the kinases. We selected the phosphosites with known regulatory status in PhosphositePlus or with unknown status but with a functional score higher than 0.4 (1,534 of 4,247 kinase-mapping phosphosites). The functional score was calculated using the *FunscoR* R package (evocellnet.github.io/funscoR/). The scores range from 0-1 and reflect the functional consequence of the phosphosites (Ochoa et al. 2020). The kinase activity inference method was the one sample z-test as described above.

TF activities were estimated using the VIPER algorithm (Alvarez et al. 2016), using log2FC as gene level statistics (see Data pre-processing and normalization - transcriptomic). VIPER was run with a minimum limit of regulon size of 5, and using all provided gene level statistics as a background (eset.filter = FALSE). VIPER returned a normalized enrichment score for 292 TFs across 1,187 cancer samples.

Benchmark of the kinase targets

We validated the kinase activities calculated from the different sources of kinase targets (database, text-mining, *in vivo* and *in vitro*) using a MS-based phosphoproteomic dataset reporting the relative phosphorylation changes of 52,814 phosphosites in 103 human perturbation-dependent conditions (Ochoa et al. 2016; Hernandez-Armenta et al. 2017). This data includes a gold standard dataset composed of 184 kinase-condition pairs where kinase regulation is expected to occur.

The z-test-based absolute kinase activity scores estimated from the different kinase substrate sources were used as classifiers of kinase regulation. Given the imbalance between the positive (gold standard) and negative (kinase-condition pairs with unknown regulation) classes, we generated 100 random sets of negative cases with the size of the positive set. The predictive skill of each classifier was evaluated by the mean area under the receiver operating characteristic curves (AUROCs). As a control, we replicated the 100 random sets of negative and positive pairs (53 pairs each) along the different lists of kinase-substrates. The ROC curves and corresponding AUCs were calculated using the *prediction* and *performance* functions from the *ROCR* R package.

Genetic associations with the kinase and TF activities

The effects of mutations on the kinase and TF activities were assessed by associating the activity of a given protein with the mutational status of the same protein or other proteins it might interact throughout the cellular regulatory networks. First, we built a binary mutation matrix M where the index M_{ij} corresponds to 1 if the sample i has a mutation in gene j and 0 otherwise. To do that, we selected the mutations classified as frameshift and in frame Indels, missense, nonsense and stop codon loss. Given the proteins X and Y, the association between the activity of Y (Y_{act}) and the mutational status of X (X_{mut}) was assessed across samples by fitting a linear model that took into account possible confounding effects:

Equation 4.2: $Y_{act} = \beta_0 + \beta_1 Study + \beta_2 X_{mut} + \varepsilon$

where Y_{act} represents the activity of protein Y, β_0 the intercept, β_1 the regression coefficient for the covariate experimental study, β_2 the regression coefficient for the mutational status of X and ε the noise term. This model was applied to assess the effect of X_{mut} on the activity of the same protein ($X_{act} \sim X_{mut}$) and on the activity of other proteins ($Y_{act} \sim X_{mut}$). The P-

values from the coefficients of X_{mut} (β_2) were calculated using the *t*-statistic over a Student's *t*-distribution and adjusted for false discovery rate (FDR) using the Benjamini-Hochberg method. The linear models and respective statistics were calculated using the *lm* and *p.adjust* R functions.

The associations were performed with the genes mutated in more than 20 samples and with the protein activities estimated in at least 10 samples. An association between a pair $Y_{act} \sim X_{mut}$ or $X_{act} \sim X_{mut}$ was performed if X_{mut} was mutated in at least 5 of all the samples in the pair. Regarding the $Y_{act} \sim X_{mut}$ associations, we tested 520,938 pairs between 208 kinases and 3,590 genes and 1,048,216 pairs between 292 TFs and 3,590 genes. In relation to the $X_{act} \sim X_{mut}$ associations, we tested 40 pairs and 64 pairs with the kinases and TFs, respectively.

Projection of the protein activities in low-dimensional spaces

We reduced the dimensionality of the kinase and TF activity matrices using the PCA and UMAP methods. Given the sparseness of the kinase activity matrix, we imputed the missing values using the *missForest* function from the *missForest* R package. Prior to that, we selected the kinases (columns) with activity measures in at least 60% of the samples and the samples (rows) with measures in at least 80% of the kinases. The imputed kinase activity matrix contained 90 kinases across 727 samples. The PCA analysis was performed using the *prcomp* R function (scale. = T, center = T) and the UMAP analysis using the *umap* function from the *umap* R package (with default parameters).

When correlating the kinase activities with the UMAP projections (Pearson correlation coefficient), we excluded redundant kinases based on the degree of shared substrates. We first performed a hierarchical clustering analysis (*hclust* R function, agglomeration method = "complete") using the Jaccard Index (JI) of shared substrates between kinases as distance measure (1-JI). Then, the kinase dendrogram was cut at a specific level (height = 0.85) to identify clusters of non-redundant kinases. We only kept one kinase per cluster (with the largest amount of substrates), reducing the number of kinases from 304 to 208.

Correlation of kinase pairs

We obtained kinase-kinase regulation pairs from the OmniPath database (omnipathdb.org/interactions). We selected the interactions reported as directed, activating (stimulating relationships) and consensual along the resources (databases). Then, we correlated the activity of the kinase-kinase pairs along the samples using the Spearman's rank correlation coefficient. The kinase pairs were stratified by the number of databases in which the interaction was found as a way of ascertaining the relevance of the interactions.

Strong correlations might be due to the amount of shared substrates between kinase pairs and not because of co-regulation events. To control for this technical limitation, we repeated the correlation analysis using a set of non-redundant kinases. This set was obtained by performing a hierarchical clustering analysis (*hclust* R function, agglomeration method = "complete") using the degree of shared substrates between kinases as distance measure (1 - Jaccard Index of shared substrates). The dendrogram tree was cut at a height cutoff of 0.8. Just one kinase was kept per cluster (with the highest number of substrates). Using this approach, we reduced the number of kinases from 304 to 231.

Associations between the activities of kinases and transcription factors

For a given protein pair K and T, where K is a kinase and T is a transcription factor, we tested whether the changes in the activity of kinase K are linearly associated with changes in the activity of the transcription factor T. To do that, we fitted a linear model to predict the activity of transcription factor T (T_{act}) using the activity of kinase K (K_{act}), while adjusting for possible confounding effects:

Equation 4.3: $T_{act} = \beta_0 + \beta_1 Study + \beta_2 K_{act} + \varepsilon$

where T_{act} represents the activity of the transcription factor T, β_0 the intercept, β_1 the regression coefficient for the covariate experimental study, β_2 the regression coefficient for the activity of kinase K and ε the noise term. The P-values from the coefficients of K_{act} (β_2) were calculated using the *t*-statistic over a Student's *t*-distribution and adjusted for false discovery rate (FDR) using the Benjamini-Hochberg method. The linear models and respective statistics were calculated using the *lm* and *p.adjust* R functions. Using this model we tested 26,280 kinase-TF associations between 90 kinases and 292 TFs.

Enrichment of the protein association pairs in the STRING network

The genetic and the kinase-TF associations were tested for enrichment in the STRING protein-protein interactions network using Fisher's exact tests (*fisher.test* R function, alternative = “greater”). The human network (version 11.0) was downloaded from the STRING database (string-db.org) as a list of protein-protein interactions with the corresponding combined scores. The scores range from 150 to 999 and represent the confidence of the respective interactions. We filtered the network using a minimum score of 850 to select the most confident protein-protein interactions. Fisher's exact tests were performed by overlapping the protein association pairs with the STRING network across increasing -log10 adjusted P-values. The backgrounds corresponded to all the protein association pairs linearly modelled.

Kinase activity changes between tumours and perturbations

To study the differences of kinase signalling between tumours and perturbation-dependent conditions, we estimated the activity of kinases across an extended panel of perturbations with phosphoproteomic measurements (Ochoa et al. 2016). This dataset is composed of 76,379 phosphosites across 439 perturbations. Next, we calculated the percentage of tumour samples and perturbations each kinase was regulated in, using an absolute kinase activity cutoff of 1.75 as previously used (Ochoa et al. 2016). We kept the kinases regulated in at least 1 tumour or perturbation. In order to find the kinases preferentially regulated in the tumours and in the perturbations - tumour or perturbation-specific kinases - we fitted a linear model between the percentage of kinase regulation in the tumours and in the perturbations, as independent and dependent variables, respectively. The most deviating kinases from the regression line were considered to be differentially regulated. These kinases were found by converting the residuals of the linear model to z-scores: while the kinases with a residual z-score > 2 were classified as tumour-specific, the kinases with a residual z-score < -2 were classified as perturbation-specific. This process was performed across all cancer samples and by tissue type. The linear models and respective residuals were calculated using the *lm* and *residuals* R functions. The residuals were standardized to z-scores using the *scale* R function.

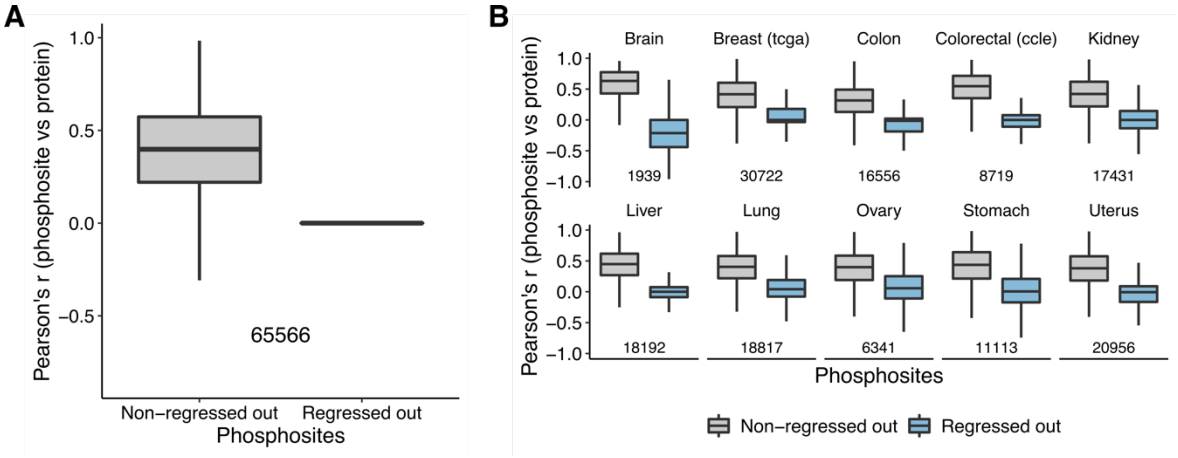
Survival analysis

In order to construct Kaplan-Meier (KM) survival curves, cancer samples were stratified based on their TF and kinase activity scores (AS). For each kinase and TF, we classified the samples as: inactive if AS < -1.75; active if AS > 1.75; neutral if -1.75 < AS < 1.75. The 1.75 activity cutoff was chosen based on a previous publication (Ochoa et al. 2016). We estimated the KM survival curves by protein and tissue. We tested if the differences on the activities of a given TF on a given tissue were associated with the probability of survival across time if: more than 10 deaths occurred and more than 10 samples were classified as active and inactive. Given the lower number of activation/inactivation events for the kinases, we tested the kinase-tissue pairs with more than 5 samples classified as active or inactive and with more than 5 deaths. These filters resulted in 1,025 tests for the TFs (274 TFs and 5 tissues) and 195 tests for the kinases (81 kinases and 7 tissues). The survival distributions of the cancer sample groups were compared using log-rank tests with the *survdiff* function from the *survival* R package. The P-values were adjusted for FDR using the Benjamini-Hochberg (BH) procedure (*p.adjust* R function). The KM curves were plotted using the *ggsurvplot* function from the *survminer* R package.

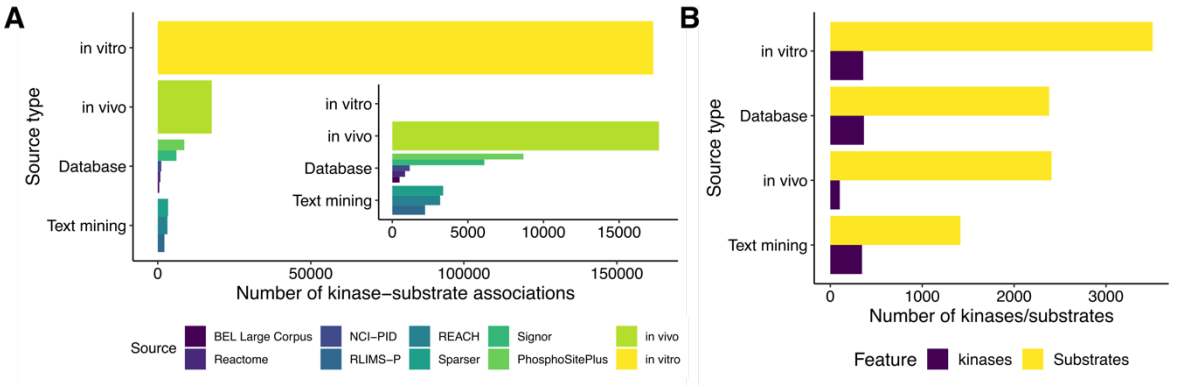
To account for confounding covariates, we performed a multivariate statistical analysis using Cox proportional-hazards regression models. The hazard function was fitted using the protein activity scores as a continuous predictor, adjusted for age, gender and the genotype (1 if mutated and 0 otherwise) of 28 recurrently mutated genes in our atlas (at least 100 mutations). Such models were applied to the protein-tissue pairs described above. We extracted the hazard ratios of the protein activity coefficients and corresponding 95% confidence intervals and P-values from the Cox models. The BH-corrected P-values were calculated using the *p.adjust* R function. The Cox regression models were fitted using the *coxph* function from the *survival* R package.

4.6. Supplementary materials

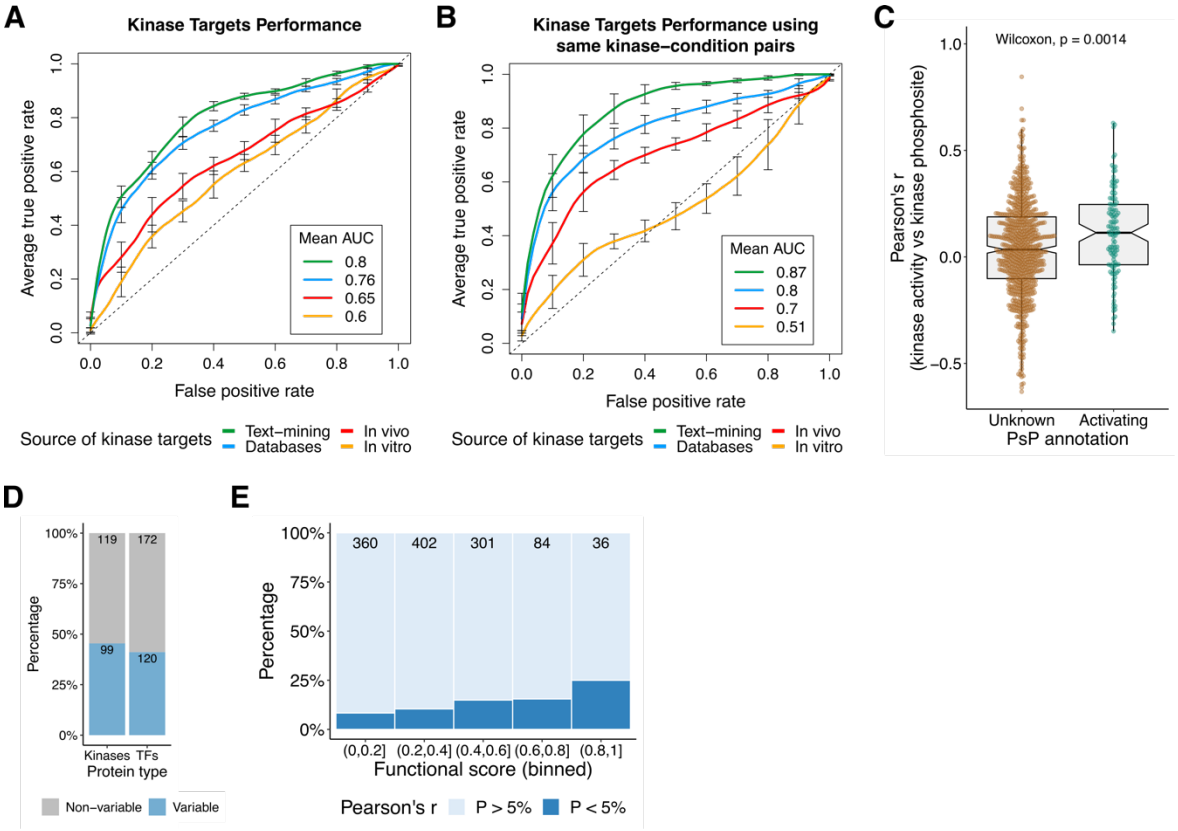
4.6.1. Figures



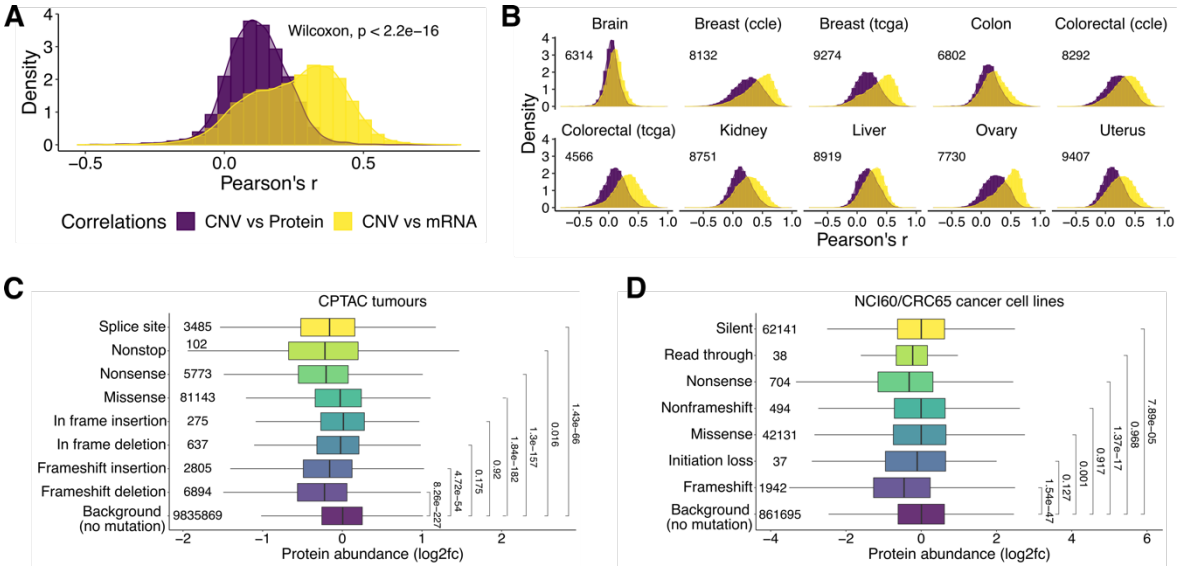
Supplementary figure 4.1. Pearson's correlation between phosphorylation levels and corresponding protein abundances. (A) Distribution of the correlations between protein abundances and phosphorylation changes for protein-phosphosite pairs (number of pairs beneath the boxplots) across all cancer samples. Given the sparseness of the (phospho)proteomics data, we selected the protein-phosphosite pairs with protein/phosphorylation measures in at least 1% (n > 10) of the total cancer samples. Left: non-regressed-out phosphorylation data. Right: protein regressed-out phosphorylation data (**Methods**). **(B)** Representation of the same data as (A) by cancer dataset. Correlations were calculated for those protein-phosphosite pairs with protein/phosphorylation measures in at least 10% (n > 5) of the samples of each dataset. A small amount of correlation between phosphosites and proteins remain as the regression was done across all of the dataset.



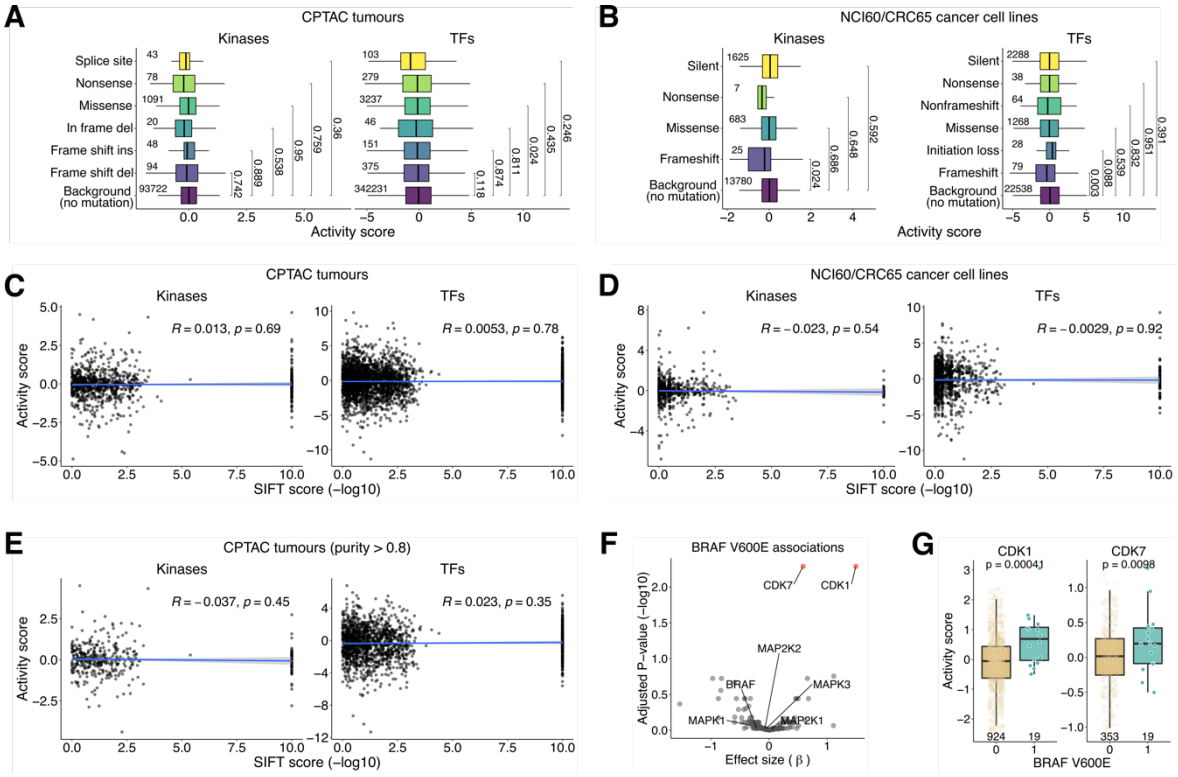
Supplementary figure 4.2. Lists of kinase-substrate associations compiled in this study. (A) Number of kinase-substrate associations by source type. **(B)** Number of kinases and substrates by source type.



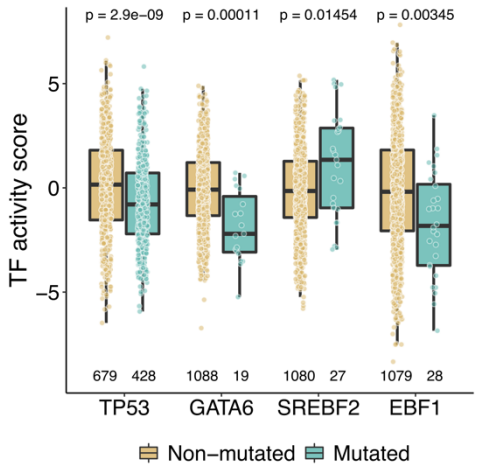
Supplementary figure 4.3. Validation of kinase-substrate sources and kinase activity estimates in the cancer samples. (A) Receiver operating characteristic (ROC) curves demonstrating the predictive performance of the Z-test-based kinase activities across different sources of kinase-substrate interactions. As positives, we used a set of 184 kinase-condition pairs where regulation is expected to occur, while as negatives we generated 100 random sets of the same size as the positive set. Curves display the average of 100 ROC curves and vertical bars the standard deviation of the true positive rate at multiple points of false positive rate. The average area under the ROC curve (AUC) is shown for each kinase-substrate list. The averaged ROC curves and corresponding AUCs demonstrate the discriminative power of each kinase-substrate list. **(B)** In contrast to the analysis shown in (A), here we replicated the 100 sets of negative (53) and positive (53) regulatory pairs along the different lists of kinase substrates. **(C)** Related to the main **Figure 4.1C**. Kinase activities were re-estimated in cancer samples after removing the kinase auto-regulatory phosphosites from the kinase targets. The boxplots show the distribution of the Pearson's correlation between kinase activities and phosphosite quantifications that mapped to the same kinase, with (n = 118) and without (n = 743) annotation (activating) in PhosphoSitePlus. **(D)** Fraction of kinases and TFs classified as highly variable across the tumour samples. Kinases and TFs with absolute activity measures higher than 1.75 and 3.89 (96.7th percentiles), respectively, in at least 5% of the samples were classified as highly variable. **(E)** Percentage of phosphosites in TFs significantly and not significantly correlated with the corresponding TF activities, stratified by their functional score. Analysis based on 1,183 phosphosites mapping to 178 TFs (n > 10).



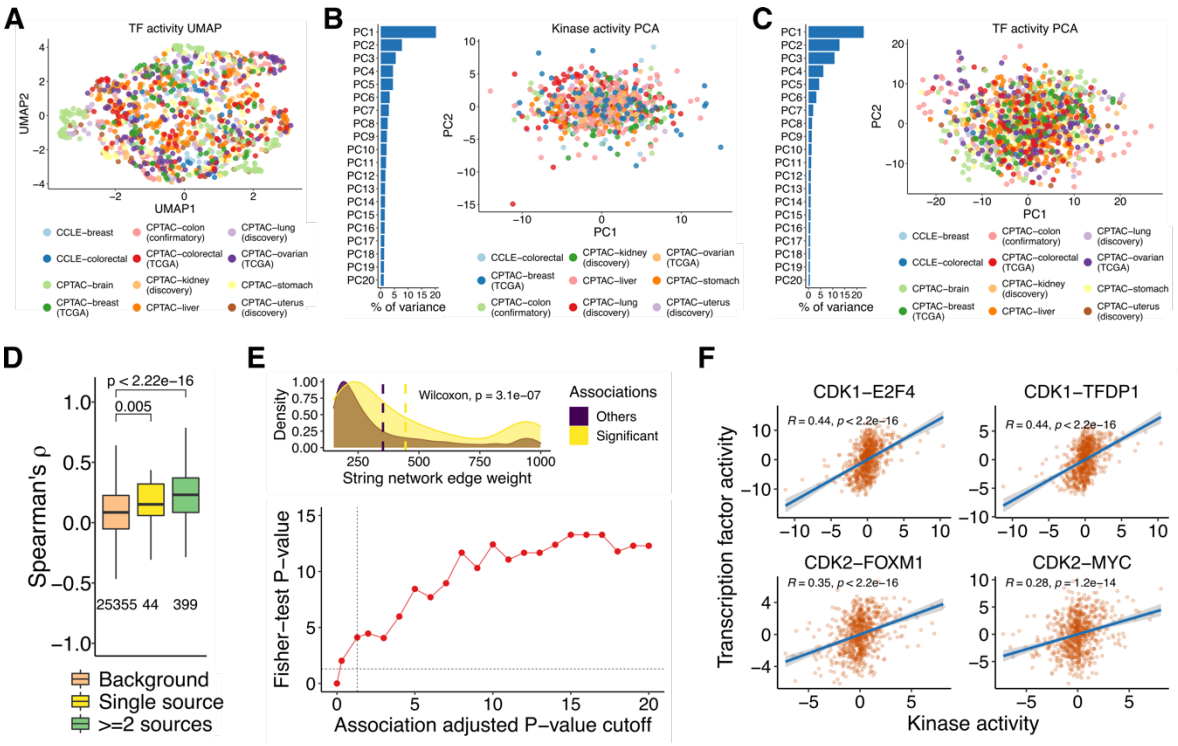
Supplementary figure 4.4. Effects of genomic alterations on protein abundances. (A) Comparison of the distribution of the correlations (Pearson's r) between the CNV levels (GISTIC2) and the mRNA and protein abundances (log2 fold-changes). Correlations were calculated for those genes with CNV, mRNA and protein quantifications in at least 10 samples (11,624 genes). The experimental batch was regressed-out from the mRNA and protein quantification data before computing the correlations (**Methods**). **(B)** Same as (A) by tissue and experimental study. P-values < 2.2e-16 in all cases (Wilcoxon rank sum test). The number of genes is indicated in the plot. **(C)** Protein abundance distribution between mutation types from the CPTAC tumours. One sample may have multiple mutations in the same protein. Therefore, we selected the sample-protein pairs that were exclusive of each mutation type to prevent the cases where different mutations in the same protein and sample have the same protein abundance. The outliers (defined as the data points beyond Q1-1.5*IQR and Q3+1.5*IQR, where Q1 and Q3 are the first and third quartiles and IQR is the interquartile range) were removed from the distributions for representation purposes. The number of protein quantifications (including outliers) is shown at the left of each boxplot. The P-values from a two-sample T-test comparing each distribution with the background (no mutation) are shown at the right. All data points (including outliers) were used to calculate the P-values. **(D)** Same as (C) for the cancer cell lines from the NCI60 and CRC65 panels.



Supplementary figure 4.5. Effects of genomic alterations on protein activities. (A) Distribution of kinase and TF activities between mutation types from the CPTAC tumours. Only the sample-protein pairs that were specific of each mutation type were selected to prevent the cases where different mutations in the same protein and sample have the same protein activity. The outliers (defined as the data points beyond Q1-1.5*IQR and Q3+1.5*IQR, where Q1 and Q3 are the first and third quartiles and IQR is the interquartile range) were removed from the distributions for representation purposes. The number of protein activity quantifications (including outliers) is shown at the left of each boxplot. The P-values from a two-sample T-test comparing each distribution with the background (no mutation) are shown at the right. All data points (including outliers) were used to calculate the P-values. **(B)** Same as (A) for the cancer cell lines from the NCI60 and CRC65 panels. **(C)** Scatterplots between the -log10 SIFT score (x-axis) of missense mutations and the activity of kinases and TFs (y-axis) from the CPTAC tumours. The linear regression line and the Pearson correlation coefficient, with the respective P-value, are shown. Cases where the same sample had multiple missense mutations in the same gene were removed to prevent the assignment of the same protein activity to different SIFT scores. **(D)** Same as (C) for the cancer cell lines from the NCI60 and CRC65 panels. **(E)** Same as (C) for the CPTAC tumours with higher purity (greater than 0.8). The purity score provides information about the degree of immune infiltration and was calculated from the gene expression data using the ESTIMATE algorithm. **(F)** Volcano plot showing the associations between the $BRAF^{V600E}$ mutation and the activity of kinases. The x-axis contains the mutation coefficient (effect size) and the y-axis the adjusted P-values. Highlighted are kinases from the MAPK/ERK signaling pathway and CDK1/7 (significantly associated with the $BRAF^{V600E}$ mutation). **(G)** Differential activity of the CDK1 and CDK7 kinases between samples with and without $BRAF^{V600E}$ mutation. The x-axis separates the cancer samples by mutation status (1 if mutated and 0 otherwise) and the y-axis contains the kinase activities. The outliers were removed from the distributions for representation purposes. The number of quantifications (including outliers) are shown beneath each boxplot. A P-value from a Wilcoxon rank sum test comparing both distributions is shown. All data points (including outliers) were used to calculate the P-values.

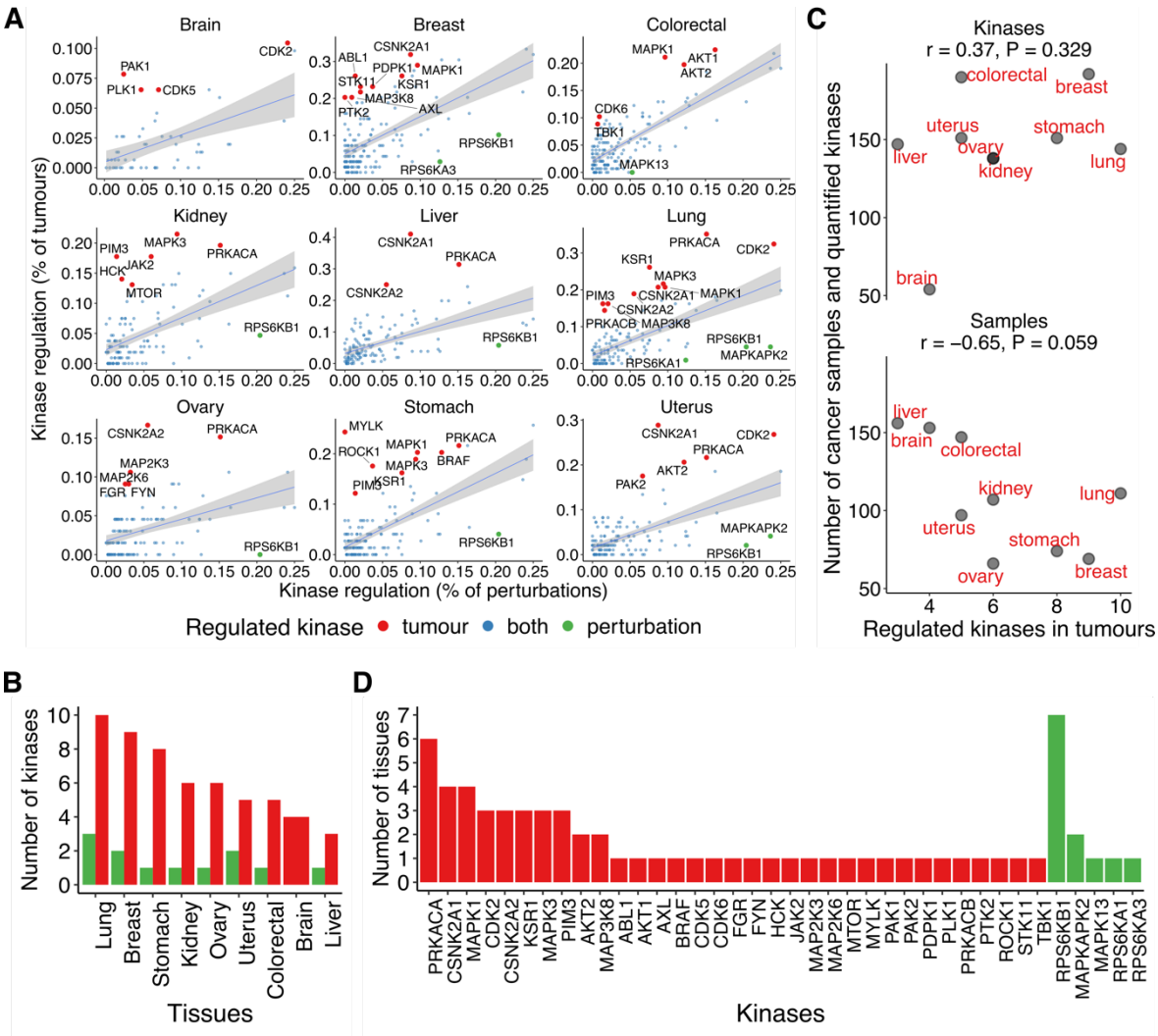


Supplementary figure 4.6. Examples of associations between the mutational status of TFs and their activities. Related to the main **Figure 4.2B**. The x-axis represents the TFs and the y-axis the activities. The colours stratify the samples by their mutational status in the respective TFs. The number of quantifications is shown beneath each boxplot. The P-values from Wilcoxon rank sum tests comparing both distributions are shown.

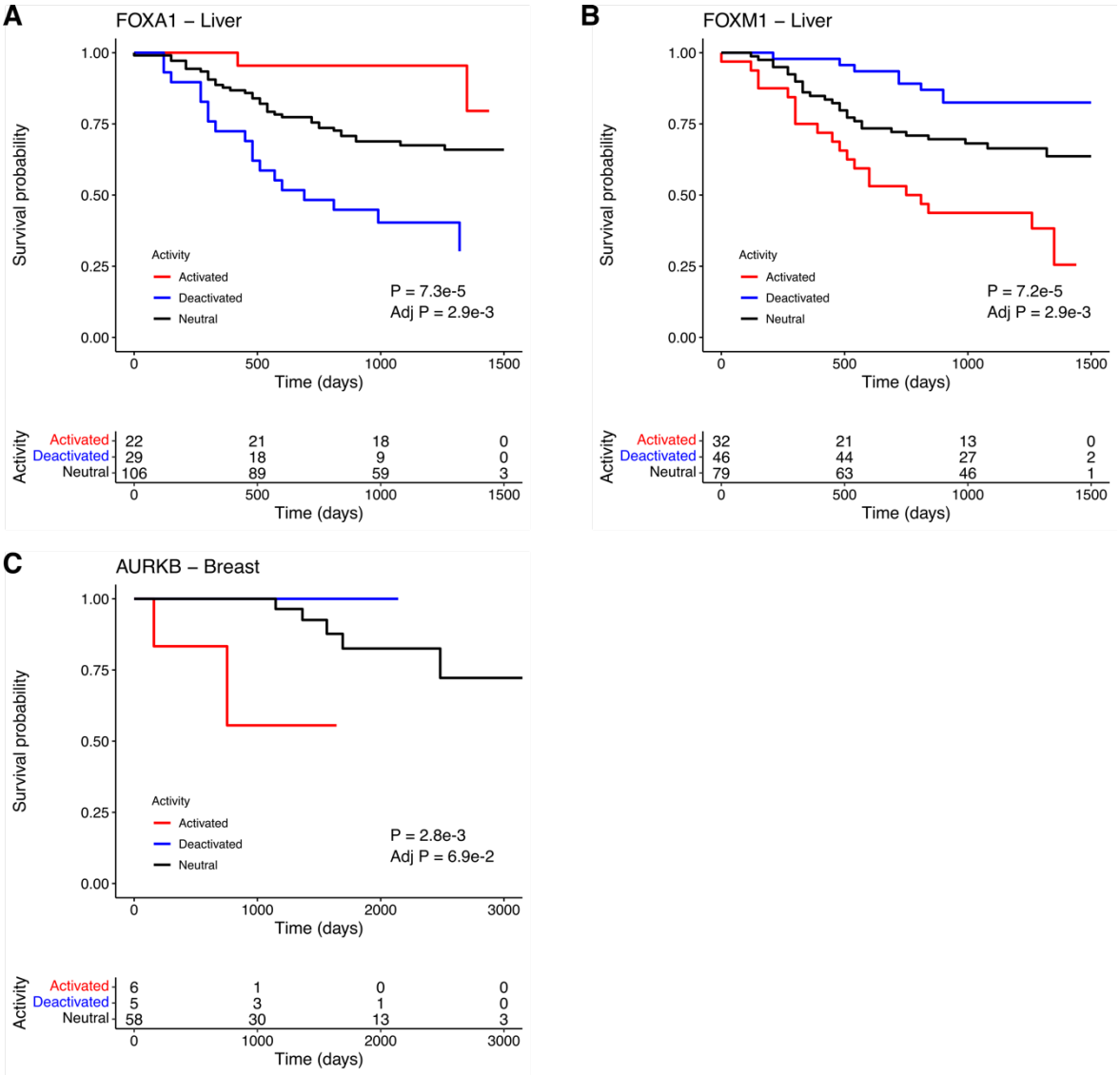


Supplementary figure 4.7. Projection of protein activities in low-dimensional spaces and kinase-TF associations. (A) UMAP projection of the TF activity matrix (TFs as variables). The samples are coloured by experimental study. **(B)** PCA of the kinase activities. The barplots indicate the percentage of total variance explained by the first 20 principal components (PCs) (out of 90 PCs). The scatter plots illustrate the samples projected along the PC1 and PC2. The samples are coloured by experimental study. **(C)** Same as (B) for the TFs. The barplot contains 20 of 292 PCs. **(D)** Related to the main **Figure 4.3D**. Correlations between the activities of non-redundant kinases with co-regulatory relationships. The co-regulatory interactions were obtained from OmniPath (activating and consensual interactions along the sources) and catalogued as present in a

single source or in at least two different sources. The background corresponds to kinase pairs for which co-regulation is not known. The distributions were compared to the background using Wilcoxon rank sum tests. **(E) Top panel.** String network edge weight distributions between the significant and non-significant kinase-TF associations (224 and 7527 pairs). The significant associations were selected with an FDR < 5% and an absolute effect size > 0.5. **(E) Bottom panel.** Enrichment of the kinase-TF associations in the string network (edge weights > 850). The y-axis shows the Fisher-test P-values (-log10) and the x-axis the adjusted P-value cutoffs (-log10) that were used to select the associations. **(F)** Scatter plots of the kinase-TF associations highlighted in the main **Figure 4.3E**.



Supplementary figure 4.8. Kinase activity regulation in tumours and perturbed human conditions. (A) Related to the main **Figure 4.3F**. Linear regression models between the percentage of kinase regulation in the perturbed conditions (x-axis) and in the tumour samples (y-axis) by tissue type. **(B)** Number of kinases classified as regulated in the tumours (red) and in the conditions (green) in each tissue. **(C)** Correlation between the number of regulated kinases in tumours (x-axis) and the number of quantified kinases and samples (y-axis) across tissues. The Pearson's r and respective P-value are shown. **(D)** Number of tissues where the kinases were identified as regulated in the tumours (red) or in the conditions (green). The kinases are mutually exclusive between them (no kinase found as regulated in the tumours and in the conditions).



Supplementary figure 4.9. The activities of FOXA1/FOXM1 and AURKB are associated with the overall survival of liver and breast cancer patients. Related to the main **Figure 4.4A, Figure 4.4B**. KM survival plots for **(A)** FOXA1 (inactive = 29, neutral = 106, active = 22) and **(B)** FOXM1 (46, 79, 32) in liver cancer and **(C)** AURKB (5, 58, 6) in breast cancer. The tables beneath each plot contain the number of individuals at risk across time. The log-rank P-values are shown in the plots.

4.6.2. Tables

Supplementary table 4.1. Kinase and TF activities estimated from the molecular data.

Supplementary table 4.2. Significant correlations between the protein activities and the corresponding CNV, RNA, protein and phosphorylation levels.

Supplementary table 4.3. Genetic associations with protein activities.

Supplementary table 4.4. Associations between kinase and TF activities.

Supplementary table 4.5. Protein activities significantly associated with the overall survival of the cancer patients.

All supplementary tables can be consulted using the following DOI:
doi.org/10.1101/2021.06.09.447741

5. Conclusions and Future Perspectives

Genomic instability is a known hallmark of cancer and is the result of somatic mutations, including gene copy-number alterations and dynamic changes in chromosome number and structure (Hanahan and Weinberg 2011). These changes lead to variation in the same cancer clone and intratumoral genetic heterogeneity (Iacobuzio-Donahue, Litchfield, and Swanton 2020). Moreover, the recent developments on cancer genomics have led to the conclusion that molecular heterogeneity prevails between different tumour subtypes within a tumour type. Consequently, each cancer patient may have a different response to classical treatments, i.e., chemotherapy and radiotherapy. Precision oncology is based on the premise that matching the oncogenic cancer targets with therapeutic agents that were engineered based on the status of the targets will improve cancer treatment (Rodriguez et al. 2021; Colomer et al. 2020). Major cancer genomics projects such as TCGA and ICGC (see subchapter **1.2** for details) have paved the road for precision oncology. These projects have defined cancer molecular subtypes that may enable treatment of patients with significantly more precision. However, a gap remains in our ability to associate genomic variations with cancer phenotypes. In fact, clinical strategies are known to be limited by mutational profiling alone (Le Tourneau et al. 2015; Saad et al. 2017). Moreover, transcriptome characterization is not enough to assess the functional consequence of most cancer mutations. It has been shown that RNA expression levels are often poor predictors of the corresponding protein levels (Gonçalves et al. 2017), which are the main targets of most anti-cancer drugs. Therefore, the study of cancer proteomes, including proteins and PTMs, will be essential to narrow the gap between the cancer genotype and phenotype. It is expected that multi-omics cancer studies and cancer proteogenomics in particular will offer new therapeutic avenues for precision oncology in the future (Rodriguez et al. 2021). In chapter **3** I demonstrated that up to 42% of CNV changes are attenuated at the protein level (see subchapter **3.3.1** for more details). Together with related findings from CPTAC studies (Mertins et al. 2016; B. Zhang et al. 2014; Vasaikar et al. 2019; Gao et al. 2019), these results may help to prioritize genomic aberrations that potentially act as both oncogenic drivers and drug targets.

Transcriptomic and phospho(proteomic) profiling at single-cell resolution will help to dissect the genetic and signalling heterogeneity that underlie the phenotypic diversity of cells within a tumour. So far, bulk transcriptomic and phospho(proteomic) studies of signalling pathways do not account for cell-to-cell variability. At the single-cell level, the activities of signalling proteins (i.e., kinases and TFs) and whole pathways can be highly variable depending on the genomic background and tumour microenvironment. In chapter **4** I discuss the lack of correlation between loss-of-function cancer mutations and kinase and TF activities (see subchapter **4.3.3** for more details). These results may be due to many sources of genetic (i.e., mutations) and non-genetic (i.e., epigenetic factors, stochasticity of

biomolecules and environmental stimuli) heterogeneity in individual cancer cells. It is expected that novel methods such as single-cell RNA-seq and single-cell proteomics by MS will allow to profile signalling networks cell-by-cell. These techniques may help to uncover the signalling consequences of intratumoral genetic and non-genetic heterogeneity, to quantify the variation of cell-to-cell signalling networks, and to evaluate downstream phenotypic effects induced by modulation of signalling pathways (Lun and Bodenmiller 2020; J. Fan, Slowikowski, and Zhang 2020; Budnik et al. 2018). In precision oncology, single-cell multi-omics tumour profiling may anticipate cancer drug resistance by resolving adaptive signalling responses and guide the clinicians through novel combinations of drug therapies (Wei et al. 2016).

Worldwide, men develop more cancers in non-reproductive tissues than women (Siegel, Miller, and Jemal 2018). In fact, the predominance of cancer in men has been observed across all races and ages (A. D. Wagner et al. 2019). Moreover, men also tend to have a worse prognosis than women. The predisposition of men to cancer is probably the consequence of a complex interaction between sex-specific biology and lifestyle. Sex-specific biology includes genetic differences on the sex chromosomes and the effects of sex hormones on cells (see subchapter **1.3.1.1** for more details). Lifestyle factors include dietary habits and risky behaviours such as smoking and alcohol consumption (Allen et al. 2016; McCartney et al. 2011). However, even after adjusting for lifestyle factors, the male bias in cancer incidence persists (Edgren et al. 2012), suggesting the existence of gender intrinsic molecular factors of predisposition/protection to cancer. In chapter **2** I conducted a deep gender-differential gene expression and co-expression network analysis to unravel the sex-biased cancer transcriptome in stomach and thyroid, two tissue types with unbalanced cancer incidences. Strikingly, I found fewer SBGs in tumours than in normal tissues, suggesting that gene expression differences between genders are diluted after tumorigenesis in these tissues (see subchapter **2.3.2** for more details). Similarly, Yuan and colleagues found very few SBGs in a set of tissues classified as the weak sex-effect group (Yuan et al. 2016). So far, most studies on gender molecular differences have focused on gene expression (J. Ma, Malladi, and Beck 2016; Lopes-Ramos et al. 2020; Aguet et al. 2020; Yuan et al. 2016). Given the limitation of gene expression to explain complex phenotypes and diseases such as cancer (Yansheng Liu, Beyer, and Aebersold 2016), phospho(proteomic) profiling should be considered in the future to study the molecular basis of sex disparities in cancer. In addition, beyond the characterization of gene/protein expression differences between genders, future studies should also follow other lines of research. For instance, the inference of protein activities from transcriptomics and phosphoproteomics data could provide mechanistic descriptions of sex-biased signalling in cancer (Dugourd et al. 2021). Lastly, future clinical trials should consider more often the

biology of sex differences to evaluate gender disparities in drug safety and efficacy (Mauvais-Jarvis et al. 2020; Buoncervello et al. 2017).

Multi-omics cancer datasets will be useful not only to study cancer onset and treatment, but also to investigate important cellular mechanisms of protein homeostasis and maintenance. For instance, in chapter **3** I described that some protein complex members act as rate-limiting for the assembly of the complex, whereby one subunit controls the abundance level of the interacting partners (see subchapter **3.3.2** for more details). Moreover, the degree of protein attenuation was correlated with the fraction of residues at interfaces. The usefulness of cancer proteogenomics datasets to study cell biology has been exemplified by other works. Ryan et al. found in breast cancer that mutation of one protein complex subunit was often associated with a collateral reduction in protein expression of other complex members (Ryan et al. 2017). The collateral loss was mainly evident at the proteomic level, suggesting post-transcriptional control. The contribution of multiple mechanisms to protein-level gene dosage buffering remains to be studied, including the control of protein translation rates by microRNAs and RNA-binding proteins and protein degradation.

Looking forward, additional opportunities in precision oncology may also involve the integration of multi-omics data with histopathological images through machine learning techniques. Yu Fu et al. recently applied a convolutional neural network to tissue images from TCGA, extracting thousands of histopathological features across more than 10,000 individuals from 28 tumour types (Y. Fu et al. 2020). The authors found that the extracted features enabled tissue classification and were associated with genetic aberrations, transcriptomic signatures and prognostic information. In the future, artificial intelligence may be routinely used in the clinic to aid tumour detection and prognosis (Esteva et al. 2017; Ciompi et al. 2017; Kooi et al. 2017; Dhungel, Carneiro, and Bradley 2017).

Precision oncology has already transformed the way cancer patients are managed and treated. The number of druggable tumour-specific molecular biomarkers has increased substantially in the past decade, which significantly improved patient survival in several cancer types (Malone et al. 2020). However, the path to the routine use of genomic biomarkers in cancer clinics is still long (Boutros 2015). Proteomics and phosphoproteomics will be essential to identify and prioritize genomic alterations that can be targeted by novel drugs (Rodriguez et al. 2021). Other challenges include the education and engagement of clinicians and patients about the benefits of precision medicine, as well as the promotion of data sharing between research institutes and hospitals to maximize knowledge gain. Nevertheless, multi-omics approaches in cancer research have the potential to identify novel biomarkers that can be used to diagnose cancer earlier and faster, to improve patient survival and response to treatment, and to monitor disease relapse.

6. Bibliography

Abovich, N, L Gritz, L Tung, and M Rosbash. 1985. “Effect of RP51 Gene Dosage Alterations on Ribosome Synthesis in Saccharomyces Cerevisiae.” *Molecular and Cellular Biology* 5 (12): 3429–35. http://www.ncbi.nlm.nih.gov/pubmed/3915776.

Aebersold, Ruedi, and Matthias Mann. 2016. “Mass-Spectrometric Exploration of Proteome Structure and Function.” *Nature*. Nature Publishing Group. https://doi.org/10.1038/nature19949.

Agrawal, Nishant, Rehan Akbani, B. Arman Aksoy, Adrian Ally, Harindra Arachchi, Sylvia L. Asa, J. Todd Auman, et al. 2014. “Integrated Genomic Characterization of Papillary Thyroid Carcinoma.” *Cell* 159 (3): 676–90. https://doi.org/10.1016/j.cell.2014.09.050.

Aguet, François, Alvaro N. Barbeira, Rodrigo Bonazzola, Andrew Brown, Stephane E. Castel, Brian Jo, Silva Kasela, et al. 2020. “The Impact of Sex on Gene Expression across Human Tissues.” *Science* 369 (6509). https://doi.org/10.1126/SCIENCE.ABA3066.

Aguet, François, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, et al. 2017. “Genetic Effects on Gene Expression across Human Tissues.” *Nature* 550 (7675): 204–13. https://doi.org/10.1038/nature24277.

Ahrens, Christian H., Erich Brunner, Ermir Qeli, Konrad Basler, and Ruedi Aebersold. 2010. “Generating and Navigating Proteome Maps Using Mass Spectrometry.” *Nature Reviews Molecular Cell Biology*. Nat Rev Mol Cell Biol. https://doi.org/10.1038/nrm2973.

Akalin, Altuna, Verdan Franke, Bora Uyar, and Jonathan Ronen. 2021. *Computational Genomics with R*. 1st ed. CRC Press Taylor & Francis Group.

Akbani, Rehan, Karl Friedrich Becker, Neil Carragher, Ted Goldstein, Leanne De Koning, Ulrike Korf, Lance Liotta, et al. 2014. “Realizing the Promise of Reverse Phase Protein Arrays for Clinical, Translational, and Basic Research: A Workshop Report the RPPA (Reverse Phase Protein Array) Society.” *Molecular and Cellular Proteomics* 13 (7): 1625–43. https://doi.org/10.1074/mcp.O113.034918.

Akbani, Rehan, Patrick Kwok Shing Ng, Henrica M.J. Werner, Maria Shahmoradgoli, Fan Zhang, Zhenlin Ju, Wenbin Liu, et al. 2014. “A Pan-Cancer Proteomic Perspective on the Cancer Genome Atlas.” *Nature Communications* 5 (May). https://doi.org/10.1038/ncomms4887.

Alexandrov, Ludmil B., Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, et al. 2016. “Mutational Signatures Associated with Tobacco Smoking in Human Cancer.” *Science* 354 (6312): 618–22. https://doi.org/10.1126/science.aag0299.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. “The Repertoire of Mutational Signatures in Human

Cancer." *Nature* 578 (7793): 94–101. https://doi.org/10.1038/s41586-020-1943-3.

Alexandrov, Ludmil B., and Michael R. Stratton. 2014. "Mutational Signatures: The Patterns of Somatic Mutations Hidden in Cancer Genomes." *Current Opinion in Genetics and Development*. Curr Opin Genet Dev. https://doi.org/10.1016/j.gde.2013.11.014.

Alfaro, Javier A., Ankit Sinha, Thomas Kislinger, and Paul C. Boutros. 2014. "Onco-Proteogenomics: Cancer Proteomics Joins Forces with Genomics." *Nature Methods*. Nature Publishing Group. https://doi.org/10.1038/nmeth.3138.

Ali, Imran, Johan Högberg, Jui Hua Hsieh, Scott Auerbach, Anna Korhonen, Ulla Stenius, and Ilona Silins. 2016. "Gender Differences in Cancer Susceptibility: Role of Oxidative Stress." *Carcinogenesis* 37 (10): 985–92. https://doi.org/10.1093/carcin/bgw076.

Allen, Alicia M., Taneisha S. Scheuermann, Nicole Nollen, Dorothy Hatsukami, and Jasjit S. Ahluwalia. 2016. "Gender Differences in Smoking Behavior and Dependence Motives among Daily and Nondaily Smokers." *Nicotine and Tobacco Research* 18 (6): 1408–13. https://doi.org/10.1093/ntr/ntv138.

Ally, Adrian, Miruna Balasundaram, Rebecca Carlsen, Eric Chuah, Amanda Clarke, Noreen Dhalla, Robert A. Holt, et al. 2017. "Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma." *Cell* 169 (7): 1327-1341.e23. https://doi.org/10.1016/j.cell.2017.05.046.

Altelaar, A. F.Maarten, Javier Munoz, and Albert J.R. Heck. 2013. "Next-Generation Proteomics: Towards an Integrative View of Proteome Dynamics." *Nature Reviews Genetics*. Nat Rev Genet. https://doi.org/10.1038/nrg3356.

Alvarez, Mariano J., Yao Shen, Federico M. Giorgi, Alexander Lachmann, B. Belinda Ding, B. Hilda Ye, and Andrea Califano. 2016. "Functional Characterization of Somatic Mutations in Cancer Using Network-Based Inference of Protein Activity." *Nature Genetics* 48 (8): 838–47. https://doi.org/10.1038/ng.3593.

Amanchy, Ramars, Jun Zhong, Henrik Molina, Raghothama Chaerkady, Akiko Iwahori, Dario Eluan Kalume, Mads Grønborg, Jos Joore, Leslie Cope, and Akhilesh Pandey. 2008. "Identification of C-Src Tyrosine Kinase Substrates Using Mass Spectrometry and Peptide Microarrays." *Journal of Proteome Research* 7 (9): 3900–3910. https://doi.org/10.1021/pr800198w.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. "HTSeq-A Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics* 31 (2): 166–69. https://doi.org/10.1093/bioinformatics/btu638.

Andersson, Robin, and Albin Sandelin. 2020. "Determinants of Enhancer and Promoter Activities of Regulatory Elements." *Nature Reviews Genetics*. Nature Research. https://doi.org/10.1038/s41576-019-0173-8.

Aran, Dvir, Roman Camarda, Justin Odegaard, Hyojung Paik, Boris Oskotsky, Gregor

Krings, Andrei Goga, Marina Sirota, and Atul J. Butte. 2017. “Comprehensive Analysis of Normal Adjacent to Tumor Transcriptomes.” *Nature Communications* 8 (1). https://doi.org/10.1038/s41467-017-01027-z.

Argelaguet, Ricard, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni, and Oliver Stegle. 2020. “MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data.” *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-02015-1.

Argelaguet, Ricard, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. 2018. “Multi-Omics Factor Analysis—a Framework for Unsupervised Integration of Multi-omics Data Sets.” *Molecular Systems Biology* 14 (6). https://doi.org/10.15252/msb.20178124.

Armstrong, Bruce K., and Anne Kricker. 2001. “The Epidemiology of UV Induced Skin Cancer.” *Journal of Photochemistry and Photobiology B: Biology* 63 (1–3): 8–18. https://doi.org/10.1016/S1011-1344(01)00198-1.

Arshad, Osama A., Vincent Danna, Vladislav A. Petyuk, Paul D. Piehowski, Tao Liu, Karin D. Rodland, and Jason E. McDermott. 2019. “An Integrative Analysis of Tumor Proteomic and Phosphoproteomic Profiles to Examine the Relationships between Kinase Activity and Phosphorylation.” *Molecular and Cellular Proteomics* 18 (8): S26–36. https://doi.org/10.1074/mcp.RA119.001540.

Aslam, Bilal, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. 2017. “Proteomics: Technologies and Their Applications.” *Journal of Chromatographic Science*. Oxford University Press. https://doi.org/10.1093/chromsci/bmw167.

Auwera, Geraldine A. Van der, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.” *Current Protocols in Bioinformatics* 43 (SUPL.43). https://doi.org/10.1002/0471250953.bi1110s43.

Bach, Karsten, Sara Pensa, Marta Grzelak, James Hadfield, David J. Adams, John C. Marioni, and Walid T. Khaled. 2017. “Differentiation Dynamics of Mammary Epithelial Cells Revealed by Single-Cell RNA Sequencing.” *Nature Communications* 8 (1). https://doi.org/10.1038/s41467-017-02001-5.

Bachman, John, Benjamin Gyori, and Peter Sorger. 2019. “Assembling a Phosphoproteomic Knowledge Base Using ProtMapper to Normalize Phosphosite Information from Databases and Text Mining.” *BioRxiv*, November, 822668. https://doi.org/10.1101/822668.

Bailey, Christopher M., Steve M.M. Sweet, Debbie L. Cunningham, Martin Zeller, John K.

Heath, and Helen J. Cooper. 2009. “SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra.” *Journal of Proteome Research* 8 (4): 1965–71. https://doi.org/10.1021/pr800917p.

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, et al. 2018. “Comprehensive Characterization of Cancer Driver Genes and Mutations.” *Cell* 173 (2): 371-385.e18. https://doi.org/10.1016/j.cell.2018.02.060.

Bantscheff, Marcus, Simone Lemeer, Mikhail M. Savitski, and Bernhard Kuster. 2012. “Quantitative Mass Spectrometry in Proteomics: Critical Review Update from 2007 to the Present.” *Analytical and Bioanalytical Chemistry*. Anal Bioanal Chem. https://doi.org/10.1007/s00216-012-6203-4.

Banu, Sakhila K., P. Govindarajulu, and Michael M. Aruldhas. 2002. “Testosterone and Estradiol Differentially Regulate TSH-Induced Thyrocyte Proliferation in Immature and Adult Rats.” *Steroids* 67 (7): 573–79. https://doi.org/10.1016/S0039-128X(02)00008-9.

Barabási, Albert László, and Eric Bonabeau. 2003. “Scale-Free Networks.” *Scientific American* 288 (5): 60–69. https://doi.org/10.1038/scientificamerican0503-60.

Baralle, Francisco E., and Jimena Giudice. 2017. “Alternative Splicing as a Regulator of Development and Tissue Identity.” *Nature Reviews Molecular Cell Biology*. Nature Publishing Group. https://doi.org/10.1038/nrm.2017.27.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. “The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity.” *Nature* 483 (7391): 603–7. https://doi.org/10.1038/nature11003.

Bass, Adam J., Vesteinn Thorsson, Ilya Shmulevich, Sheila M. Reynolds, Michael Miller, Brady Bernard, Toshinori Hinoue, et al. 2014. “Comprehensive Molecular Characterization of Gastric Adenocarcinoma.” *Nature* 513 (7517): 202–9. https://doi.org/10.1038/nature13480.

Battle, Alexis, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. 2015. “Impact of Regulatory Variation from RNA to Protein.” *Science* 347 (6222): 664–67. https://doi.org/10.1126/science.1260793.

Beausoleil, Sean A., Judit Villén, Scott A. Gerber, John Rush, and Steven P. Gygi. 2006. “A Probability-Based Approach for High-Throughput Protein Phosphorylation Analysis and Site Localization.” *Nature Biotechnology* 24 (10): 1285–92. https://doi.org/10.1038/nbt1240.

Bell, D., A. Berchuck, M. Birrer, J. Chien, D. W. Cramer, F. Dao, R. Dhir, et al. 2011. “Integrated Genomic Analyses of Ovarian Carcinoma.” *Nature* 474 (7353): 609–15. https://doi.org/10.1038/nature10166.

Bellott, Daniel W., Jennifer F. Hughes, Helen Skaletsky, Laura G. Brown, Tatyana Pyntikova, Ting Jan Cho, Natalia Koutseva, et al. 2014. “Mammalian y Chromosomes Retain Widely Expressed Dosage-Sensitive Regulators.” *Nature* 508 (7497): 494–99. https://doi.org/10.1038/nature13206.

Beltrao, Pedro, Véronique Albanèse, Lillian R. Kenner, Danielle L. Swaney, Alma Burlingame, Judit Villén, Wendell A. Lim, James S. Fraser, Judith Frydman, and Nevan J. Krogan. 2012. “Systematic Functional Prioritization of Protein Posttranslational Modifications.” *Cell* 150 (2): 413–25. https://doi.org/10.1016/j.cell.2012.05.036.

Ben-Levy, Rachel, Hugh F. Paterson, Christopher J. Marshall, and Yosef Yarden. 1994. “A Single Autophosphorylation Site Confers Oncogenicity to the Neu/ErbB-2 Receptor and Enables Coupling to the MAP Kinase Pathway.” *EMBO Journal* 13 (14): 3302–11. https://doi.org/10.1002/j.1460-2075.1994.tb06632.x.

Berger, Michael F., and Elaine R. Mardis. 2018. “The Emerging Clinical Relevance of Genomics in Cancer Medicine.” *Nature Reviews Clinical Oncology*. Nature Publishing Group. https://doi.org/10.1038/s41571-018-0002-6.

Beroukhim, Rameen, Gad Getz, Leia Nghiemphu, Jordi Barretina, Teli Hsueh, David Linhart, Igor Vivanco, et al. 2007. “Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (50): 20007–12. https://doi.org/10.1073/pnas.0710052104.

Beroukhim, Rameen, Craig H. Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, et al. 2010. “The Landscape of Somatic Copy-Number Alteration across Human Cancers.” *Nature* 463 (7283): 899–905. https://doi.org/10.1038/nature08822.

Betts, Matthew J., Oliver Wichmann, Mathias Utz, Timon Andre, Evangelia Petsalaki, Pablo Minguez, Luca Parca, et al. 2017. “Systematic Identification of Phosphorylation-Mediated Protein Interaction Switches.” *PLoS Computational Biology* 13 (3). https://doi.org/10.1371/journal.pcbi.1005462.

Bhagwat, Anand S., and Christopher R. Vakoc. 2015. “Targeting Transcription Factors in Cancer.” *Trends in Cancer*. Cell Press. https://doi.org/10.1016/j.trecan.2015.07.001.

Bhullar, Khushwant S., Naiara Orrego Lagarón, Eileen M. McGowan, Indu Parmar, Amitabh Jha, Basil P. Hubbard, and H. P.Vasantha Rupasinghe. 2018. “Kinase-Targeted Cancer Therapies: Progress, Challenges and Future Directions.” *Molecular Cancer*. BioMed Central Ltd. https://doi.org/10.1186/s12943-018-0804-2.

Bi, Junfeng, Taka Aki Ichu, Ciro Zanca, Huijun Yang, Wei Zhang, Yuchao Gu, Sudhir Chowdhry, et al. 2019. “Oncogene Amplification in Growth Factor Signaling Pathways Renders Cancers Dependent on Membrane Lipid Remodeling.” *Cell Metabolism* 30

(3): 525-538.e8. https://doi.org/10.1016/j.cmet.2019.06.014.

Blume-Jensen, Peter, and Tony Hunter. 2001. “Oncogenic Kinase Signalling.” *Nature*. Nature. https://doi.org/10.1038/35077225.

Bosc, D. G., E. Slominski, C. Sichler, and D. W. Litchfield. 1995. “Phosphorylation of Casein Kinase II by P34(Cdc2). Identification of Phosphorylation Sites Using Phosphorylation Site Mutants in Vitro.” *Journal of Biological Chemistry* 270 (43): 25872–78. https://doi.org/10.1074/jbc.270.43.25872.

Bouhaddou, Mehdi, Danish Memon, Bjoern Meyer, Kris M. White, Veronica V. Rezelj, Miguel Correa Marrero, Benjamin J. Polacco, et al. 2020. “The Global Phosphorylation Landscape of SARS-CoV-2 Infection.” *Cell* 182 (3): 685-712.e19. https://doi.org/10.1016/j.cell.2020.06.034.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber. 2010. “Independent Filtering Increases Detection Power for High-Throughput Experiments.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (21): 9546–51. https://doi.org/10.1073/pnas.0914005107.

Boutros, Paul C. 2015. “The Path to Routine Use of Genomic Biomarkers in the Cancer Clinic.” *Genome Research*. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/gr.191114.115.

Bretones, Gabriel, M. Dolores Delgado, and Javier León. 2015. “Myc and Cell Cycle Control.” *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. Elsevier B.V. https://doi.org/10.1016/j.bbagrm.2014.03.013.

Broido, Anna D., and Aaron Clauset. 2019. “Scale-Free Networks Are Rare.” *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-08746-5.

Bucciantini, Monica, Elisa Giannoni, Fabrizio Chiti, Fabiana Baroni, Niccolò Taddei, Giampietro Ramponi, Christopher M. Dobson, and Massimo Stefani. 2002. “Inherent Toxicity of Aggregates Implies a Common Mechanism for Protein Misfolding Diseases.” *Nature* 416 (6880): 507–11. https://doi.org/10.1038/416507a.

Budnik, Bogdan, Ezra Levy, Guillaume Harmange, and Nikolai Slavov. 2018. “SCoPE-MS: Mass Spectrometry of Single Mammalian Cells Quantifies Proteome Heterogeneity during Cell Differentiation.” *Genome Biology* 19 (1). https://doi.org/10.1186/s13059-018-1547-5.

Buoncervello, Maria, Matteo Marconi, Alessandra Carè, Paola Piscopo, Walter Malorni, and Paola Matarrese. 2017. “Preclinical Models in the Study of Sex Differences.” *Clinical Science* 131 (6): 449–69. https://doi.org/10.1042/CS20160847.

Bushweller, John H. 2019. “Targeting Transcription Factors in Cancer — from Undruggable to Reality.” *Nature Reviews Cancer*. Nature Publishing Group. https://doi.org/10.1038/s41568-019-0196-7.

Calabrese, Claudia, Natalie R. Davidson, Deniz Demircioglu, Nuno A. Fonseca, Yao He, André Kahles, Kjong Van Lehmann, et al. 2020. “Genomic Basis for RNA Alterations in Cancer.” *Nature* 578 (7793): 129–36. https://doi.org/10.1038/s41586-020-1970-0.

Campbell, Peter J., Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, et al. 2020. “Pan-Cancer Analysis of Whole Genomes.” *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Carrel, Laura, and Huntington F. Willard. 2005. “X-Inactivation Profile Reveals Extensive Variability in X-Linked Gene Expression in Females.” *Nature* 434 (7031): 400–404. https://doi.org/10.1038/nature03479.

Casado, Pedro, Juan Carlos Rodriguez-Prados, Sabina C. Cosulich, Sylvie Guichard, Bart Vanhaesebroeck, Simon Joel, and Pedro R. Cutillas. 2013. “Kinase-Substrate Enrichment Analysis Provides Insights into the Heterogeneity of Signaling Pathway Activation in Leukemia Cells.” *Science Signaling*. Sci Signal. https://doi.org/10.1126/scisignal.2003573.

Chen, Fengju, Darshan S. Chandrashekar, Sooryanarayana Varambally, and Chad J. Creighton. 2019. “Pan-Cancer Molecular Subtypes Revealed by Mass-Spectrometry-Based Proteomic Characterization of More than 500 Human Cancers.” *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-13528-0.

Chick, Joel M., Deepak Kolippakkam, David P. Nusinow, Bo Zhai, Ramin Rad, Edward L. Huttlin, and Steven P. Gygi. 2015. “A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides.” *Nature Biotechnology* 33 (7): 743–49. https://doi.org/10.1038/nbt.3267.

Chick, Joel M., Steven C. Munger, Petr Simecek, Edward L. Huttlin, Kwangbom Choi, Daniel M. Gatti, Narayanan Raghupathy, Karen L. Svenson, Gary A. Churchill, and Steven P. Gygi. 2016. “Defining the Consequences of Genetic Variation on a Proteome-Wide Scale.” *Nature* 534 (7608): 500–505. https://doi.org/10.1038/nature18270.

Choudhary, Chunaram, and Matthias Mann. 2010. “Decoding Signalling Networks by Mass Spectrometry-Based Proteomics.” *Nature Reviews Molecular Cell Biology*. Nat Rev Mol Cell Biol. https://doi.org/10.1038/nrm2900.

Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. 2013. “Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples.” *Nature Biotechnology* 31 (3): 213–19. https://doi.org/10.1038/nbt.2514.

Ciompi, Francesco, Kaman Chung, Sarah J. Van Riel, Arnaud Arindra Adiyoso Setio, Paul K. Gerke, Colin Jacobs, Ernst Th Scholten, et al. 2017. “Towards Automatic Pulmonary Nodule Management in Lung Cancer Screening with Deep Learning.” *Scientific*

Reports 7 (April). https://doi.org/10.1038/srep46479.

Clark, David J., Saravana M. Dhanasekaran, Francesca Petralia, Jianbo Pan, Xiaoyu Song, Yingwei Hu, Felipe da Veiga Leprevost, et al. 2019. “Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma.” *Cell* 179 (4): 964-983.e31. https://doi.org/10.1016/j.cell.2019.10.007.

Clauser, Karl R., Peter Baker, and Alma L. Burlingame. 1999. “Role of Accurate Mass Measurement (±10 Ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching.” *Analytical Chemistry* 71 (14): 2871–82. https://doi.org/10.1021/ac9810516.

Clocchiatti, Andrea, Elisa Cora, Yosra Zhang, and G. Paolo Dotto. 2016. “Sexual Dimorphism in Cancer.” *Nature Reviews Cancer* 16 (5): 330–39. https://doi.org/10.1038/nrc.2016.30.

Colaprico, Antonio, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, et al. 2016. “TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data.” *Nucleic Acids Research* 44 (8): e71. https://doi.org/10.1093/nar/gkv1507.

Colomer, Ramon, Rebeca Mondejar, Nuria Romero-Laorden, Arantzazu Alfranca, Francisco Sanchez-Madrid, and Miguel Quintela-Fandino. 2020. “When Should We Order a next Generation Sequencing Test in a Patient with Cancer?” *EClinicalMedicine*. Lancet Publishing Group. https://doi.org/10.1016/j.eclinm.2020.100487.

Costa-Silva, Juliana, Douglas Domingues, and Fabricio Martins Lopes. 2017. “RNA-Seq Differential Expression Analysis: An Extended Review and a Software Tool.” *PLoS ONE*. Public Library of Science. https://doi.org/10.1371/journal.pone.0190152.

Cox, Jürgen, and Matthias Mann. 2008. “MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification.” *Nature Biotechnology* 26 (12): 1367–72. https://doi.org/10.1038/nbt.1511.

———. 2011. “Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology.” *Annual Review of Biochemistry* 80 (July): 273–99. https://doi.org/10.1146/annurev-biochem-061308-093216.

Cox, Jurgen, Ivan Matic, Maximiliane Hilger, Nagarjuna Nagaraj, Matthias Selbach, Jesper V. Olsen, and Matthias Mann. 2009. “A Practical Guide to the Maxquant Computational Platform for Silac-Based Quantitative Proteomics.” *Nature Protocols* 4 (5): 698–705. https://doi.org/10.1038/nprot.2009.36.

Cox, Jürgen, Nadin Neuhauser, Annette Michalski, Richard A. Scheltema, Jesper V. Olsen, and Matthias Mann. 2011. “Andromeda: A Peptide Search Engine Integrated into the

MaxQuant Environment." *Journal of Proteome Research* 10 (4): 1794–1805. https://doi.org/10.1021/pr101065j.

Creixell, Pau, Erwin M. Schoof, Craig D. Simpson, James Longden, Chad J. Miller, Hua Jane Lou, Lara Perryman, et al. 2015. "Kinome-Wide Decoding of Network-Attacking Mutations Rewiring Cancer Signaling." *Cell* 163 (1): 202–17. https://doi.org/10.1016/j.cell.2015.08.056.

Croft, David, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, et al. 2014. "The Reactome Pathway Knowledgebase." *Nucleic Acids Research* 42 (D1). https://doi.org/10.1093/nar/gkt1102.

D'Angelo, Gina, Raghothama Chaerkady, Wen Yu, Deniz Baycin Hizal, Sonja Hess, Wei Zhao, Kristen Lekstrom, et al. 2017. "Statistical Models for the Analysis of Isobaric Tags Multiplexed Quantitative Proteomics." *Journal of Proteome Research* 16 (9): 3124–36. https://doi.org/10.1021/acs.jproteome.6b01050.

Dalgliesh, Gillian L., Kyle Furge, Chris Greenman, Lina Chen, Graham Bignell, Adam Butler, Helen Davies, et al. 2010. "Systematic Sequencing of Renal Carcinoma Reveals Inactivation of Histone Modifying Genes." *Nature* 463 (7279): 360–63. https://doi.org/10.1038/nature08672.

Darnell, James E. 2002. "Transcription Factors as Targets for Cancer Therapy." *Nature Reviews Cancer*. Nat Rev Cancer. https://doi.org/10.1038/nrc906.

Davies, M. A., and Y. Samuels. 2010. "Analysis of the Genome to Personalize Therapy for Melanoma." *Oncogene*. Oncogene. https://doi.org/10.1038/onc.2010.323.

Dephoure, Noah, Sunyoung Hwang, Ciara O'Sullivan, Stacie E Dodgson, Steven P Gygi, Angelika Amon, and Eduardo M Torres. 2014. "Quantitative Proteomic Analysis Reveals Posttranslational Responses to Aneuploidy in Yeast." *ELife* 3 (July): e03023. https://doi.org/10.7554/eLife.03023.

Deschênes-Simard, Xavier, Filippos Kottakis, Sylvain Meloche, and Gerardo Ferbeyre. 2014. "ERKs in Cancer: Friends or Foes?" *Cancer Research*. Cancer Res. https://doi.org/10.1158/0008-5472.CAN-13-2381.

Dhillon, A. S., S. Hagan, O. Rath, and W. Kolch. 2007. "MAP Kinase Signalling Pathways in Cancer." *Oncogene*. Oncogene. https://doi.org/10.1038/sj.onc.1210421.

Dhungel, Neeraj, Gustavo Carneiro, and Andrew P. Bradley. 2017. "A Deep Learning Approach for the Analysis of Masses in Mammograms with Minimal User Intervention." *Medical Image Analysis* 37 (April): 114–28. https://doi.org/10.1016/j.media.2017.01.009.

Ding, Li, Matthew H. Bailey, Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L. Gibbs, et al. 2018. "Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics." *Cell* 173 (2): 305-320.e10.

https://doi.org/10.1016/j.cell.2018.03.033.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics (Oxford, England)* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dou, Yongchao, Emily A. Kawaler, Daniel Cui Zhou, Marina A. Gritsenko, Chen Huang, Lili Blumenberg, Alla Karpova, et al. 2020. “Proteogenomic Characterization of Endometrial Carcinoma.” *Cell* 180 (4): 729-748.e26. https://doi.org/10.1016/j.cell.2020.01.026.

Drake, Justin M., Nicholas A. Graham, Tanya Stoyanova, Amir Sedghi, Andrew S. Goldstein, Houjian Cai, Daniel A. Smith, et al. 2012. “Oncogene-Specific Activation of Tyrosine Kinase Networks during Prostate Cancer Progression.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (5): 1643–48. https://doi.org/10.1073/pnas.1120985109.

Drolz, A., M. Wewalka, T. Horvatits, V. Fuhrmann, B. Schneeweiss, M. Trauner, and C. Zauner. 2014. “Gender-Specific Differences in Energy Metabolism during the Initial Phase of Critical Illness.” *European Journal of Clinical Nutrition* 68 (6): 707–11. https://doi.org/10.1038/ejcn.2013.287.

Dugourd, Aurelien, Christoph Kuppe, Marco Sciacovelli, Enio Gjerga, Attila Gabor, Kristina B. Emdal, Vitor Vieira, et al. 2021. “Causal Integration of Multi-omics Data with Prior Knowledge to Generate Mechanistic Hypotheses.” *Molecular Systems Biology* 17 (1). https://doi.org/10.15252/msb.20209730.

Dugourd, Aurelien, and Julio Saez-Rodriguez. 2019. “Footprint-Based Functional Analysis of Multiomic Data.” *Current Opinion in Systems Biology*. Elsevier Ltd. https://doi.org/10.1016/j.coisb.2019.04.002.

Dunford, Andrew, David M. Weinstock, Virginia Savova, Steven E. Schumacher, John P. Cleary, Akinori Yoda, Timothy J. Sullivan, et al. 2017. “Tumor-Suppressor Genes That Escape from X-Inactivation Contribute to Cancer Sex Bias.” *Nature Genetics* 49 (1): 10–16. https://doi.org/10.1038/ng.3726.

Dunham, Ian, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, et al. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

Edgren, Gustaf, Liming Liang, Hans Olov Adami, and Ellen T. Chang. 2012. “Enigmatic Sex Disparities in Cancer Incidence.” *European Journal of Epidemiology* 27 (3): 187–96. https://doi.org/10.1007/s10654-011-9647-5.

Edwards, Nathan J., Mauricio Oberti, Ratna R. Thangudu, Shuang Cai, Peter B. McGarvey,

Shine Jacob, Subha Madhavan, and Karen A. Ketchum. 2015. “The CPTAC Data Portal: A Resource for Cancer Proteomics Research.” *Journal of Proteome Research* 14 (6): 2707–13. https://doi.org/10.1021/pr501254j.

Elias, Joshua E., and Steven P. Gygi. 2007. “Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry.” *Nature Methods* 4 (3): 207–14. https://doi.org/10.1038/nmeth1019.

Eng, Jimmy K., Ashley L. McCormack, and John R. Yates. 1994. “An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.” *Journal of the American Society for Mass Spectrometry* 5 (11): 976–89. https://doi.org/10.1016/1044-0305(94)80016-2.

Engers, R., S. Mrzyk, E. Springer, D. Fabbro, G. Weissgerber, C. D. Gerharz, and H. E. Gabbert. 2000. “Protein Kinase C in Human Renal Cell Carcinomas: Role in Invasion and Differential Isoenzyme Expression.” *British Journal of Cancer* 82 (5): 1063–69. https://doi.org/10.1054/bjoc.1999.1043.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.” *Nature* 542 (7639): 115–18. https://doi.org/10.1038/nature21056.

Fabregat, Antonio, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, et al. 2016. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 44 (D1): D481–87. https://doi.org/10.1093/nar/gkv1351.

Fan, Jean, Kamil Slowikowski, and Fan Zhang. 2020. “Single-Cell Transcriptomics in Cancer: Computational Challenges and Opportunities.” *Experimental and Molecular Medicine*. Springer Nature. https://doi.org/10.1038/s12276-020-0422-0.

Fan, Yu, Liu Xi, Daniel S.T. Hughes, Jianjun Zhang, Jianhua Zhang, P. Andrew Futreal, David A. Wheeler, and Wenyi Wang. 2016. “MuSE: Accounting for Tumor Heterogeneity Using a Sample-Specific Error Model Improves Sensitivity and Specificity in Mutation Calling from Sequencing Data.” *Genome Biology* 17 (1): 178. https://doi.org/10.1186/s13059-016-1029-6.

Felicetti, Francesco, Maria Graziella Catalano, and Nicoletta Fortunati. 2017. “Thyroid Autoimmunity and Cancer.” In *Frontiers of Hormone Research*, 48:97–109. https://doi.org/10.1159/000452909.

Ferreira, Pedro G., Pedro Jares, Daniel Rico, Gonzalo Gómez-López, Alejandra Martínez-Trillos, Neus Villamor, Simone Ecker, et al. 2014. “Transcriptome Characterization by RNA Sequencing Identifies a Major Molecular and Clinical Subdivision in Chronic Lymphocytic Leukemia.” *Genome Research* 24 (2): 212–26.

https://doi.org/10.1101/gr.152132.112.

Frejno, Martin, Chen Meng, Benjamin Ruprecht, Thomas Oellerich, Sebastian Scheich, Karin Kleigrewe, Enken Drecoll, et al. 2020. “Proteome Activity Landscapes of Tumor Cell Lines Determine Drug Responses.” *Nature Communications* 11 (1). https://doi.org/10.1038/s41467-020-17336-9.

Fu, Dragony, Jennifer A. Calvo, and Leona D. Samson. 2012. “Balancing Repair and Tolerance of DNA Damage Caused by Alkylating Agents.” *Nature Reviews Cancer*. Nat Rev Cancer. https://doi.org/10.1038/nrc3185.

Fu, Yu, Alexander W. Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R. Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. 2020. “Pan-Cancer Computational Histopathology Reveals Mutations, Tumor Composition and Prognosis.” *Nature Cancer* 1 (8): 800–810. https://doi.org/10.1038/s43018-020-0085-8.

Gabitova, Linara, Andrey Gorin, and Igor Astsaturov. 2014. “Molecular Pathways: Sterols and Receptor Signaling in Cancer.” *Clinical Cancer Research* 20 (1): 28–34. https://doi.org/10.1158/1078-0432.CCR-13-0122.

Gao, Qiang, Hongwen Zhu, Liangqing Dong, Weiwei Shi, Ran Chen, Zhijian Song, Chen Huang, et al. 2019. “Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma.” *Cell* 179 (2): 561-577.e22. https://doi.org/10.1016/j.cell.2019.08.052.

Garcia-Alonso, Luz, Christian H. Holland, Mahmoud M. Ibrahim, Denes Turei, and Julio Saez-Rodriguez. 2019. “Benchmark and Integration of Resources for the Estimation of Human Transcription Factor Activities.” *Genome Research* 29 (8): 1363–75. https://doi.org/10.1101/gr.240663.118.

Garcia-Alonso, Luz, Francesco Iorio, Angela Matchan, Nuno Fonseca, Patricia Jaaks, Gareth Peat, Miguel Pignatelli, et al. 2018. “Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer.” *Cancer Research* 78 (3): 769–80. https://doi.org/10.1158/0008-5472.CAN-17-1679.

Garraway, Levi A., and Eric S. Lander. 2013. “Lessons from the Cancer Genome.” *Cell* 153 (1): 17–37. https://doi.org/10.1016/j.cell.2013.03.002.

Geeleher, Paul, Aritro Nath, Fan Wang, Zhenyu Zhang, Alvaro N. Barbeira, Jessica Fessler, Robert L. Grossman, Cathal Seoighe, and R. Stephanie Huang. 2018. “Cancer Expression Quantitative Trait Loci (EQTLs) Can Be Determined from Heterogeneous Tumor Gene Expression Data by Modeling Variation in Tumor Purity.” *Genome Biology* 19 (1). https://doi.org/10.1186/s13059-018-1507-0.

Geiger, Tamar, Juergen Cox, and Matthias Mann. 2010. “Proteomic Changes Resulting from Gene Copy Number Variations in Cancer Cells.” *PLoS Genetics* 6 (9).

https://doi.org/10.1371/journal.pgen.1001090.

Gershoni, Moran, and Shmuel Pietrokovski. 2017. “The Landscape of Sex-Differential Transcriptome and Its Consequent Selection in Human Adults.” *BMC Biology* 15 (1). https://doi.org/10.1186/s12915-017-0352-z.

Gholami, Amin Moghaddas, Hannes Hahne, Zhixiang Wu, Florian Johann Auer, Chen Meng, Mathias Wilhelm, and Bernhard Kuster. 2013. “Global Proteome Analysis of the NCI-60 Cell Line Panel.” *Cell Reports* 4 (3): 609–20. https://doi.org/10.1016/j.celrep.2013.07.018.

Gilbert, Ethel S. 2009. “Ionising Radiation and Cancer Risks: What Have We Learned from Epidemiology?” *International Journal of Radiation Biology* 85 (6): 467–82. https://doi.org/10.1080/09553000902883836.

Gillette, Michael A., Shankha Satpathy, Song Cao, Saravana M. Dhanasekaran, Suhas V. Vasaikar, Karsten Krug, Francesca Petralia, et al. 2020. “Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma.” *Cell* 182 (1): 200-225.e35. https://doi.org/10.1016/j.cell.2020.06.013.

Giurgiu, Madalina, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. 2019. “CORUM: The Comprehensive Resource of Mammalian Protein Complexes - 2019.” *Nucleic Acids Research* 47 (D1): D559–63. https://doi.org/10.1093/nar/gky973.

Gonçalves, Emanuel, Athanassios Fragoulis, Luz Garcia-Alonso, Thorsten Cramer, Julio Saez-Rodriguez, and Pedro Beltrao. 2017. “Widespread Post-Transcriptional Attenuation of Genomic Copy-Number Variation in Cancer.” *Cell Systems* 5 (4): 386-398.e4. https://doi.org/10.1016/j.cels.2017.08.013.

Gong, Jing, Shufang Mei, Chunjie Liu, Yu Xiang, Youqiong Ye, Zhao Zhang, Jing Feng, et al. 2018. “PancanQTL: Systematic Identification of Cis -EQTLs and Trans -EQTLs in 33 Cancer Types.” *Nucleic Acids Research* 46 (D1): D971–76. https://doi.org/10.1093/nar/gkx861.

Gonzalez-Perez, Abel, Jordi Deu-Pons, and Nuria Lopez-Bigas. 2012. “Improving the Prediction of the Functional Impact of Cancer Mutations by Baseline Tolerance Transformation.” *Genome Medicine* 4 (11). https://doi.org/10.1186/gm390.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg.2016.49.

Gou, Qian, Xin Gong, Jianhua Jin, Juanjuan Shi, and Yongzhong Hou. 2017. “Peroxisome Proliferator-Activated Receptors (PPARs) Are Potential Drug Targets for Cancer Therapy.” *Oncotarget*. Impact Journals LLC. https://doi.org/10.18632/oncotarget.19610.

Greenberg, Maxim V.C., and Deborah Bourc'his. 2019. "The Diverse Roles of DNA Methylation in Mammalian Development and Disease." *Nature Reviews Molecular Cell Biology*. Nature Publishing Group. https://doi.org/10.1038/s41580-019-0159-6.

Greenfield, Andy, Laura Carrel, David Pennisi, Christophe Philippe, Nandita Quaderi, Pamela Siggers, Kirsten Steiner, et al. 1998. "The UTX Gene Escapes X Inactivation in Mice and Humans." *Human Molecular Genetics* 7 (4): 737–42. https://doi.org/10.1093/hmg/7.4.737.

Guha, Udayan, Raghothama Chaerkady, Arivusudar Marimuthu, A. Scott Patterson, Manoj K. Kashyap, H. C. Harsha, Mitsuo Sato, et al. 2008. "Comparisons of Tyrosine Phosphorylated Proteins in Cells Expressing Lung Cancer-Specific Alleles of EGFR and KRAS." *Proceedings of the National Academy of Sciences of the United States of America* 105 (37): 14112–17. https://doi.org/10.1073/pnas.0806158105.

Guo, Ailan, Judit Villén, Jon Kornhauser, Kimberly A. Lee, Matthew P. Stokes, Klarisa Rikova, Anthony Possemato, et al. 2008. "Signaling Networks Assembled by Oncogenic EGFR and C-Met." *Proceedings of the National Academy of Sciences of the United States of America* 105 (2): 692–97. https://doi.org/10.1073/pnas.0707270105.

Haaften, Gijs Van, Gillian L. Dalgliesh, Helen Davies, Lina Chen, Graham Bignell, Chris Greenman, Sarah Edkins, et al. 2009. "Somatic Mutations of the Histone H3K27 Demethylase Gene UTX in Human Cancer." *Nature Genetics* 41 (5): 521–23. https://doi.org/10.1038/ng.349.

Halazonetis, Thanos D., Vassilis G. Gorgoulis, and Jiri Bartek. 2008. "An Oncogene-Induced DNA Damage Model for Cancer Development." *Science*. Science. https://doi.org/10.1126/science.1140735.

Han, Yonghua, Bin Ma, and Kaizhong Zhang. 2005. "Spider: Software for Protein Identification from Sequence Tags with de Novo Sequencing Error." *Journal of Bioinformatics and Computational Biology* 3 (3): 697–716. https://doi.org/10.1142/S0219720005001247.

Hanahan, Douglas, and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The next Generation." *Cell* 144 (5): 646–74. https://doi.org/10.1016/j.cell.2011.02.013.

Harari, D., and Y. Yarden. 2000. "Molecular Mechanisms Underlying ErbB2/HER2 Action in Breast Cancer." *Oncogene*. Oncogene. https://doi.org/10.1038/sj.onc.1203973.

Hartwell, Leland H, Leroy Hood, Michael L Goldberg, Ann E Reynolds, and Lee M Silver. 2011. "Genetics: From Genes to Genomes." In , 4th ed., 256, 259–64. The McGraw-Hill Companies.

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. 2017. "Multi-Omics Approaches to Disease." *Genome Biology*. BioMed Central Ltd. https://doi.org/10.1186/s13059-017-

1215-1.

Hausen, Harald Zur. 1991. “Viruses in Human Cancers.” *Science* 254 (5035): 1167–73. https://doi.org/10.1126/science.1659743.

He, Shujiao, Junyi Zhang, Wan Zhang, Fengsheng Chen, and Rongcheng Luo. 2017. “FOXA1 Inhibits Hepatocellular Carcinoma Progression by Suppressing PIK3R1 Expression in Male Patients.” *Journal of Experimental and Clinical Cancer Research* 36 (1). https://doi.org/10.1186/s13046-017-0646-6.

Hernandez-Armenta, Claudia, David Ochoa, Emanuel Gonçalves, Julio Saez-Rodriguez, and Pedro Beltrao. 2017. “Benchmarking Substrate-Based Kinase Activity Inference Using Phosphoproteomic Data.” *Bioinformatics* 33 (12): 1845–51. https://doi.org/10.1093/bioinformatics/btx082.

Higgins, Reneé, Joshua M. Gendron, Lisa Rising, Raymond Mak, Kristofor Webb, Stephen E. Kaiser, Nathan Zuzow, et al. 2015. “The Unfolded Protein Response Triggers Site-Specific Regulatory Ubiquitylation of 40S Ribosomal Proteins.” *Molecular Cell* 59 (1): 35–49. https://doi.org/10.1016/j.molcel.2015.04.026.

Hijazi, Maruan, Ryan Smith, Vinothini Rajeeve, Conrad Bessant, and Pedro R. Cutillas. 2020. “Reconstructing Kinase Network Topologies from Phosphoproteomics Data Reveals Cancer-Associated Rewiring.” *Nature Biotechnology* 38 (4): 493–502. https://doi.org/10.1038/s41587-019-0391-9.

Hiltemann, Saskia, Guido Jenster, Jan Trapman, Peter Van Der Spek, and Andrew Stubbs. 2015. “Discriminating Somatic and Germline Mutations in Tumor DNA Samples without Matching Normals.” *Genome Research* 25 (9): 1382–90. https://doi.org/10.1101/gr.183053.114.

Hoadley, Katherine A., Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, et al. 2018. “Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.” *Cell* 173 (2): 291-304.e6. https://doi.org/10.1016/j.cell.2018.03.022.

Hoeijmakers, Jan H.J. 2009. “DNA Damage, Aging, and Cancer.” *New England Journal of Medicine* 361 (15): 1475–85. https://doi.org/10.1056/nejmra0804615.

Hoffert, Jason D., and Mark A. Knepper. 2008. “Taking Aim at Shotgun Phosphoproteomics.” *Analytical Biochemistry*. Academic Press Inc. https://doi.org/10.1016/j.ab.2007.11.023.

Holoch, Daniel, and Danesh Moazed. 2015. “RNA-Mediated Epigenetic Regulation of Gene Expression.” *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg3863.

Hornbeck, Peter V., Bin Zhang, Beth Murray, Jon M. Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. 2015. “PhosphoSitePlus, 2014: Mutations, PTMs and

Recalibrations." *Nucleic Acids Research* 43 (D1): D512–20. https://doi.org/10.1093/nar/gku1267.

Huang, Di, Yu Huang, Zisheng Huang, Jiefeng Weng, Shuai Zhang, and Weili Gu. 2019. "Relation of AURKB Over-Expression to Low Survival Rate in BCRA and Reversine-Modulated Aurora B Kinase in Breast Cancer Cell Lines." *Cancer Cell International* 19 (1). https://doi.org/10.1186/s12935-019-0885-z.

Huang, Kuan lin, R. Jay Mashl, Yige Wu, Deborah I. Ritter, Jiayin Wang, Clara Oh, Marta Paczkowska, et al. 2018. "Pathogenic Germline Variants in 10,389 Adult Cancers." *Cell* 173 (2): 355-370.e14. https://doi.org/10.1016/j.cell.2018.03.039.

Hutter, Sonja, Sara Bolin, Holger Weishaupt, and Fredrik J. Swartling. 2017. "Modeling and Targeting MYC Genes in Childhood Brain Tumors." *Genes*. MDPI AG. https://doi.org/10.3390/genes8040107.

Iacobuzio-Donahue, Christine A., Kevin Litchfield, and Charles Swanton. 2020. "Intratumor Heterogeneity Reflects Clinical Disease Course." *Nature Cancer* 1 (1): 3–6. https://doi.org/10.1038/s43018-019-0002-1.

Invergo, Brandon M., Borgthor Petursson, Nosheen Akhtar, David Bradley, Girolamo Giudice, Maruan Hijazi, Pedro Cutillas, Evangelia Petsalaki, and Pedro Beltrao. 2020. "Prediction of Signed Protein Kinase Regulatory Circuits." *Cell Systems* 10 (5): 384-396.e9. https://doi.org/10.1016/j.cels.2020.04.005.

Ishikawa, Koji, Koji Makanae, Shintaro Iwasaki, Nicholas T. Ingolia, and Hisao Moriya. 2017. "Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes." *PLoS Genetics* 13 (1). https://doi.org/10.1371/journal.pgen.1006554.

Itsara, Andy, Gregory M. Cooper, Carl Baker, Santhosh Girirajan, Jun Li, Devin Absher, Ronald M. Krauss, et al. 2008. "Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease." *American Journal of Human Genetics* 84 (2): 148–61. https://doi.org/10.1016/j.ajhg.2008.12.014.

J, Cox, Hein MY, Luber CA, Paron I, Nagaraj N, and Mann M. 2014. "Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ." *Molecular & Cellular Proteomics : MCP* 13 (9): 2513–26. https://doi.org/10.1074/MCP.M113.031591.

Jackson, Stephen P., and Jiri Bartek. 2009. "The DNA-Damage Response in Human Biology and Disease." *Nature*. Nature. https://doi.org/10.1038/nature08467.

Jara, Evelyn L., Natalia Muñoz-Durango, Carolina Llanos, Carlos Fardella, Pablo A. González, Susan M. Bueno, Alexis M. Kalergis, and Claudia A. Riedel. 2017. "Modulating the Function of the Immune System by Thyroid Hormones and Thyrotropin." *Immunology Letters*. Elsevier B.V.

https://doi.org/10.1016/j.imlet.2017.02.010.

Jensen, Ole N. 2006. “Interpreting the Protein Language Using Proteomics.” *Nature Reviews Molecular Cell Biology*. Nat Rev Mol Cell Biol. https://doi.org/10.1038/nrm1939.

Jeong, Jun Soo, Hyun Ki Kim, Cho Rok Lee, Seulkee Park, Jae Hyun Park, Sang Wook Kang, Jong Ju Jeong, Kee Hyun Nam, Woong Youn Chung, and Cheong Soo Park. 2012. “Coexistence of Chronic Lymphocytic Thyroiditis with Papillary Thyroid Carcinoma: Clinical Manifestation and Prognostic Outcome.” *Journal of Korean Medical Science* 27 (8): 883–89. https://doi.org/10.3346/jkms.2012.27.8.883.

Jiang, Lihua, Meng Wang, Shin Lin, Ruiqi Jian, Xiao Li, Joanne Chan, Guanlan Dong, et al. 2020. “A Quantitative Proteome Map of the Human Body.” *Cell* 183 (1): 269-283.e19. https://doi.org/10.1016/j.cell.2020.08.036.

Jiao, Yunshen, Lingyu Ding, Ming Chu, Tieshan Wang, Jiarui Kang, Xiaofan Zhao, Huanhuan Li, et al. 2017. “Effects of Cancer-Testis Antigen, TFDP3, on Cell Cycle Regulation and Its Mechanism in L-02 and HepG2 Cell Lines in Vitro.” *PLoS ONE* 12 (8). https://doi.org/10.1371/journal.pone.0182781.

Johnson, Erik C.B., Eric B. Dammer, Duc M. Duong, Lingyan Ping, Maotian Zhou, Luming Yin, Lenora A. Higginbotham, et al. 2020. “Large-Scale Proteomic Analysis of Alzheimer’s Disease Brain and Cerebrospinal Fluid Reveals Early Changes in Energy Metabolism Associated with Microglia and Astrocyte Activation.” *Nature Medicine* 26 (5): 769–80. https://doi.org/10.1038/s41591-020-0815-6.

Johnson, Sam A., and Tony Hunter. 2005. “Kinomics: Methods for Deciphering the Kinome.” *Nature Methods*. Nat Methods. https://doi.org/10.1038/nmeth731.

Jonas, Stefanie, and Elisa Izaurralde. 2015. “Towards a Molecular Understanding of MicroRNA-Mediated Gene Silencing.” *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg3965.

Kaizu, Kazunari, Hisao Moriya, and Hiroaki Kitano. 2010. “Fragilities Caused by Dosage Imbalance in Regulation of the Budding Yeast Cell Cycle.” *PLoS Genetics* 6 (4). https://doi.org/10.1371/journal.pgen.1000919.

Kalra, Mamta, Jary Mayes, Senait Assefa, Anil K. Kaul, and Rashmi Kaul. 2008. “Role of Sex Steroid Receptors in Pathobiology of Hepatocellular Carcinoma.” *World Journal of Gastroenterology*. Baishideng Publishing Group Co. https://doi.org/10.3748/wjg.14.5945.

Kammers, Kai, Robert N. Cole, Calvin Tiengwe, and Ingo Ruczinski. 2015. “Detecting Significant Changes in Protein Abundance.” *EuPA Open Proteomics* 7 (June): 11–19. https://doi.org/10.1016/j.euprot.2015.02.002.

Kanehisa, Minoru, and Susumu Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and

Genomes." *Nucleic Acids Research*. Oxford University Press. https://doi.org/10.1093/nar/28.1.27.

Karunagaran, Devarajan, Eldad Tzahar, Roger R. Beerli, Xiaomei Chen, Diana Graus-Porta, Barry J. Ratzkin, Rony Seger, Nancy E. Hynes, and Yosef Yarden. 1996. "ErbB-2 Is a Common Auxiliary Subunit of NDF and EGF Receptors: Implications for Breast Cancer." *EMBO Journal* 15 (2): 254–64. https://doi.org/10.1002/j.1460-2075.1996.tb00356.x.

Khurana, Ekta, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A. Rubin, and Mark Gerstein. 2016. "Role of Non-Coding Sequence Variants in Cancer." *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg.2015.17.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. "HISAT: A Fast Spliced Aligner with Low Memory Requirements." *Nature Methods* 12 (4): 357–60. https://doi.org/10.1038/nmeth.3317.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. 2013. "TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." *Genome Biology* 14 (4): R36. https://doi.org/10.1186/gb-2013-14-4-r36.

Kim, Kimberly H., and Charles W.M. Roberts. 2016. "Targeting EZH2 in Cancer." *Nature Medicine*. Nature Publishing Group. https://doi.org/10.1038/nm.4036.

Kim, Min Sik, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S. Manda, Raghothama Chaerkady, Anil K. Madugundu, et al. 2014. "A Draft Map of the Human Proteome." *Nature* 509 (7502): 575–81. https://doi.org/10.1038/nature13302.

Kim, Myoung Jun, Sung Kyung Choi, Seong Hwi Hong, Jung Woo Eun, Suk Woo Nam, Jeung Whan Han, and Jueng Soo You. 2018. "Oncogenic IL7R Is Downregulated by Histone Deacetylase Inhibitor in Esophageal Squamous Cell Carcinoma via Modulation of Acetylated FOXO1." *International Journal of Oncology* 53 (1): 395–403. https://doi.org/10.3892/ijo.2018.4392.

Kim, Seon Young, and David J. Volsky. 2005. "PAGE: Parametric Analysis of Gene Set Enrichment." *BMC Bioinformatics* 6 (June). https://doi.org/10.1186/1471-2105-6-144.

Kim, Woong, Eric J. Bennett, Edward L. Huttlin, Ailan Guo, Jing Li, Anthony Possemato, Mathew E. Sowa, et al. 2011. "Systematic and Quantitative Assessment of the Ubiquitin-Modified Proteome." *Molecular Cell* 44 (2): 325–40. https://doi.org/10.1016/j.molcel.2011.08.025.

Kinzler, Kenneth W., and Bert Vogelstein. 1997. "Gatekeepers and Caretakers." *Nature*. Nature Publishing Group. https://doi.org/10.1038/386761a0.

Klein, Sabra L., and Katie L. Flanagan. 2016. "Sex Differences in Immune Responses." *Nature Reviews Immunology*. Nature Publishing Group.

https://doi.org/10.1038/nri.2016.90.

Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. “Chromatin Accessibility and the Regulatory Epigenome.” *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/s41576-018-0089-8.

Klinge, Carolyn M. 2012. “MiRNAs and Estrogen Action.” *Trends in Endocrinology & Metabolism* 23 (5): 223–33. https://doi.org/10.1016/J.TEM.2012.03.002.

Koboldt, Daniel C. 2020. “Best Practices for Variant Calling in Clinical Sequencing.” *Genome Medicine*. BioMed Central Ltd. https://doi.org/10.1186/s13073-020-00791-w.

Koboldt, Daniel C., Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, et al. 2012. “Comprehensive Molecular Portraits of Human Breast Tumours.” *Nature* 490 (7418): 61–70. https://doi.org/10.1038/nature11412.

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. “VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing.” *Genome Research* 22 (3): 568–76. https://doi.org/10.1101/gr.129684.111.

Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Jason C.H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, et al. 2015. “Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation.” *Cell Stem Cell* 17 (4): 471–85. https://doi.org/10.1016/j.stem.2015.09.011.

Kooi, Thijs, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. 2017. “Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions.” *Medical Image Analysis* 35 (January): 303–12. https://doi.org/10.1016/j.media.2016.07.007.

Labonté, Benoit, Olivia Engmann, Immanuel Purushothaman, Caroline Menard, Junshi Wang, Chunfeng Tan, Joseph R. Scarpa, et al. 2017. “Sex-Specific Transcriptional Signatures in Human Depression.” *Nature Medicine* 23 (9): 1102–11. https://doi.org/10.1038/nm.4386.

Lafitte, Marie, Isabelle Moranvillier, Stéphane Garcia, Evelyne Peuchant, Juan Iovanna, Benoit Rousseau, Pierre Dubus, et al. 2013. “FGFR3 Has Tumor Suppressor Properties in Cells with Epithelial Phenotype.” *Molecular Cancer* 12 (1). https://doi.org/10.1186/1476-4598-12-83.

Langfelder, Peter, and Steve Horvath. 2008. “WGCNA: An R Package for Weighted Correlation Network Analysis.” *BMC Bioinformatics* 9: 559. https://doi.org/10.1186/1471-2105-9-559.

Langfelder, Peter, Bin Zhang, and Steve Horvath. 2008. “Defining Clusters from a

Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R." *Bioinformatics* 24 (5): 719–20. https://doi.org/10.1093/bioinformatics/btm563.

Lapek, John D., Patricia Greninger, Robert Morris, Arnaud Amzallag, Iulian Pruteanu-Malinici, Cyril H. Benes, and Wilhelm Haas. 2017. "Detection of Dysregulated Protein-Association Networks by High-Throughput Proteomics Predicts Cancer Vulnerabilities." *Nature Biotechnology* 35 (10): 983–89. https://doi.org/10.1038/nbt.3955.

Larson, David E., Christopher C. Harris, Ken Chen, Daniel C. Koboldt, Travis E. Abbott, David J. Dooling, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. 2012. "Somaticsniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data." *Bioinformatics* 28 (3): 311–17. https://doi.org/10.1093/bioinformatics/btr665.

Law, Charity W., Monther Alhamdoosh, Shian Su, Gordon K. Smyth, and Matthew E. Ritchie. 2016. "RNA-Seq Analysis Is Easy as 1-2-3 with Limma, Glimma and EdgeR." *F1000Research* 5 (0): 1408. https://doi.org/10.12688/f1000research.9005.1.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2). https://doi.org/10.1186/gb-2014-15-2-r29.

Lawrence, Robert T, Elizabeth M Perez, Daniel Hernández, Chris P Miller, Kelsey M Haas, Hanna Y Irie, Su-In Lee, C Anthony Blau, and Judit Villén. 2015. "The Proteomic Landscape of Triple-Negative Breast Cancer." *Cell Reports* 11 (4): 630–44. https://doi.org/10.1016/j.celrep.2015.03.050.

Lee, Eva Y.H.P., and William J. Muller. 2010. "Oncogenes and Tumor Suppressor Genes." *Cold Spring Harbor Perspectives in Biology*. Cold Spring Harb Perspect Biol. https://doi.org/10.1101/cshperspect.a003236.

Lee, M. L., G. G. Chen, A. C. Vlantis, G. M.K. Tse, B. C.H. Leung, and C. A. Van Hasselt. 2005. "Induction of Thyroid Papillary Carcinoma Cell Proliferation by Estrogen Is Associated with an Altered Expression of Bcl-XL." *Cancer Journal* 11 (2): 113–21. https://doi.org/10.1097/00130404-200503000-00006.

Lee, Tong Ihn, and Richard A. Young. 2013. "Transcriptional Regulation and Its Misregulation in Disease." *Cell*. Elsevier B.V. https://doi.org/10.1016/j.cell.2013.02.014.

Levin, Ellis R., and Stephen R. Hammes. 2016. "Nuclear Receptors Outside the Nucleus: Extranuclear Signalling by Steroid Receptors." *Nature Reviews Molecular Cell Biology*. Nature Publishing Group. https://doi.org/10.1038/nrm.2016.122.

Li, Bo, and Colin N Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome." *BMC Bioinformatics* 12: 323.

https://doi.org/10.1186/1471-2105-12-323.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Hong-Dong. 2018. “GTFtools: A Python Package for Analyzing Various Modes of Gene Models.” *BioRxiv*, February, 263517. https://doi.org/10.1101/263517.

Li, Jun, and Glenda Gobe. 2006. “Protein Kinase C Activation and Its Role in Kidney Disease (Review Article).” *Nephrology*. Nephrology (Carlton). https://doi.org/10.1111/j.1440-1797.2006.00673.x.

Li, Jun, Yiling Lu, Rehan Akbani, Zhenlin Ju, Paul L. Roebuck, Wenbin Liu, Ji Yeon Yang, et al. 2013. “TCPA: A Resource for Cancer Functional Proteomics Data.” *Nature Methods*. Nat Methods. https://doi.org/10.1038/nmeth.2650.

Li, Jun, and Robert Tibshirani. 2013. “Finding Consistent Patterns: A Nonparametric Approach for Identifying Differential Expression in RNA-Seq Data.” *Statistical Methods in Medical Research* 22 (5): 519–36. https://doi.org/10.1177/0962280211428386.

Li, Zhaoyu, Geetu Tuteja, Jonathan Schug, and Klaus H. Kaestner. 2012. “Foxa1 and Foxa2 Are Essential for Sexual Dimorphism in Liver Cancer.” *Cell* 148 (1–2): 72–83. https://doi.org/10.1016/j.cell.2011.11.026.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Libert, Claude, Lien Dejager, and Iris Pinheiro. 2010. “The X Chromosome in Immune Functions: When a Chromosome Makes the Difference.” *Nature Reviews. Immunology* 10 (8): 594–604. https://doi.org/10.1038/nri2815.

Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. 2011. “Molecular Signatures Database (MSigDB) 3.0.” *Bioinformatics* 27 (12): 1739–40. https://doi.org/10.1093/bioinformatics/btr260.

Lin, Che Pin, Chien Ru Liu, Chun Nin Lee, Tze Sian Chan, and H. Eugene Liu. 2010. “Targeting C-Myc as a Novel Approach for Hepatocellular Carcinoma.” *World Journal of Hepatology* 2 (1): 16–20. https://doi.org/10.4254/wjh.v2.i1.16.

Lindahl, T., and D. E. Barnes. 2000. “Repair of Endogenous DNA Damage.” In *Cold Spring Harbor Symposia on Quantitative Biology*, 65:127–33. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/sqb.2000.65.127.

Linding, Rune, Lars Juhl Jensen, Adrian Pasculescu, Marina Olhovsky, Karen Colwill, Peer Bork, Michael B. Yaffe, and Tony Pawson. 2008. “NetworKIN: A Resource for Exploring Cellular Phosphorylation Networks.” *Nucleic Acids Research* 36 (SUPPL. 1). https://doi.org/10.1093/nar/gkm902.

Lito, Piro, Neal Rosen, and David B. Solit. 2013. “Tumor Adaptation and Resistance to RAF Inhibitors.” *Nature Medicine*. Nat Med. https://doi.org/10.1038/nm.3392.

Liu, Fei, Na Chen, Ruihai Xiao, Weichao Wang, and Zhengyue Pan. 2016. “MiR-144-3p Serves as a Tumor Suppressor for Renal Cell Carcinoma and Inhibits Its Invasion and Metastasis by Targeting MAP3K8.” *Biochemical and Biophysical Research Communications* 480 (1): 87–93. https://doi.org/10.1016/j.bbrc.2016.10.004.

Liu, Jianfang, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, et al. 2018. “An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.” *Cell* 173 (2): 400-416.e11. https://doi.org/10.1016/j.cell.2018.02.052.

Liu, Yansheng, Andreas Beyer, and Ruedi Aebersold. 2016. “On the Dependency of Cellular Protein Levels on MRNA Abundance.” *Cell* 165 (3): 535–50. https://doi.org/10.1016/j.cell.2016.03.014.

Liu, Yining, Jingchun Sun, and Min Zhao. 2017. “ONGene: A Literature-Based Database for Human Oncogenes.” *Journal of Genetics and Genomics*. Institute of Genetics and Developmental Biology. https://doi.org/10.1016/j.jgg.2016.12.004.

Liu, Yusen, Edward G. Shepherd, and Leif D. Nelin. 2007. “MAPK Phosphatases - Regulating the Immune Response.” *Nature Reviews Immunology*. Nat Rev Immunol. https://doi.org/10.1038/nri2035.

Lonardo, F., E. Di Marco, C. R. King, J. H. Pierce, O. Segatto, S. A. Aaronson, and P. P. Di Fiore. 1990. “The Normal ErbB-2 Product Is an Atypical Receptor-like Tyrosine Kinase with Constitutive Activity in the Absence of Ligand.” *New Biologist* 2 (11): 992–1003. https://europepmc.org/article/med/1983208.

Lopes-Ramos, Camila M., Cho Yi Chen, Marieke L. Kuijjer, Joseph N. Paulson, Abhijeet R. Sonawane, Maud Fagny, John Platig, Kimberly Glass, John Quackenbush, and Dawn L. DeMeo. 2020. “Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues.” *Cell Reports* 31 (12). https://doi.org/10.1016/j.celrep.2020.107795.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12). https://doi.org/10.1186/s13059-014-0550-8.

Lucca, Ilaria, Tobias Klatte, Harun Fajkovic, Michela De Martino, and Shahrokh F. Shariat. 2015. “Gender Differences in Incidence and Outcomes of Urothelial and Kidney

Cancer (Nature Reviews Urology (2015) 12 (653)).” *Nature Reviews Urology*. Nature Publishing Group. https://doi.org/10.1038/nrurol.2015.257.

Lun, Xiao Kang, and Bernd Bodenmiller. 2020. “Profiling Cell Signaling Networks at Single-Cell Resolution.” *Molecular and Cellular Proteomics*. American Society for Biochemistry and Molecular Biology Inc. https://doi.org/10.1074/mcp.R119.001790.

Lundby, Alicia, Giulia Franciosa, Kristina B. Emdal, Jan C. Refsgaard, Sebastian P. Gnosa, Dorte B. Bekker-Jensen, Anna Secher, et al. 2019. “Oncogenic Mutations Rewire Signaling Pathways by Switching Protein Recruitment to Phosphotyrosine Sites.” *Cell* 179 (2): 543-560.e26. https://doi.org/10.1016/j.cell.2019.09.008.

Ma, Bin, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. 2003. “PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry.” *Rapid Communications in Mass Spectrometry* 17 (20): 2337–42. https://doi.org/10.1002/rcm.1196.

Ma, Cheng Lung, Cheng Lung Hsu, Ming Heng Wu, Chun Te Wu, Cheng Chia Wu, Jiann Jyh Lai, Yuh Shan Jou, Chun Wei Chen, Shuyuan Yeh, and Chawnshang Chang. 2008. “Androgen Receptor Is a New Potential Therapeutic Target for the Treatment of Hepatocellular Carcinoma.” *Gastroenterology* 135 (3). https://doi.org/10.1053/j.gastro.2008.05.046.

Ma, Jonathan, Sadhika Malladi, and Andrew H. Beck. 2016. “Systematic Analysis of Sex-Linked Molecular Alterations and Therapies in Cancer.” *Scientific Reports* 6 (January). https://doi.org/10.1038/srep19119.

Ma, Ze Qiang, Surendra Dasari, Matthew C. Chambers, Michael D. Litton, Scott M. Sobecki, Lisa J. Zimmerman, Patrick J. Halvey, et al. 2009. “IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering.” *Journal of Proteome Research* 8 (8): 3872–81. https://doi.org/10.1021/pr900360j.

Mackenzie, Richard W.A., and Bradley T. Elliott. 2014. “Akt/PKB Activation and Insulin Signaling: A Novel Insulin Signaling Pathway in the Treatment of Type 2 Diabetes.” *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*. Dove Medical Press Ltd. https://doi.org/10.2147/DMSO.S48260.

Macůrek, Libor, Arne Lindqvist, Dan Lim, Michael A. Lampson, Rob Klompmaker, Raimundo Freire, Christophe Clouin, Stephen S. Taylor, Michael B. Yaffe, and René H. Medema. 2008. “Polo-like Kinase-1 Is Activated by Aurora A to Promote Checkpoint Recovery.” *Nature* 455 (7209): 119–23. https://doi.org/10.1038/nature07185.

Malone, Eoghan R., Marc Oliva, Peter J.B. Sabatini, Tracy L. Stockley, and Lillian L. Siu. 2020. “Molecular Profiling for Precision Cancer Therapies.” *Genome Medicine*. BioMed Central. https://doi.org/10.1186/s13073-019-0703-1.

Mar, B. G., L. Bullinger, E. Basu, K. Schlis, L. B. Silverman, K. Döhner, and S. A. Armstrong.

2012. “Sequencing Histone-Modifying Enzymes Identifies UTX Mutations in Acute Lymphoblastic Leukemia.” *Leukemia*. Leukemia. https://doi.org/10.1038/leu.2012.56.

Mardamshina, Mariya, and Tamar Geiger. 2017. “Next-Generation Proteomics and Its Application to Clinical Breast Cancer Research.” *American Journal of Pathology*. Elsevier Inc. https://doi.org/10.1016/j.ajpath.2017.07.003.

Mardis, Elaine R. 2018. “Insights from Large-Scale Cancer Genome Sequencing.” *Annual Review of Cancer Biology*. Annual Reviews Inc. https://doi.org/10.1146/annurev-cancerbio-050216-122035.

Martin, Jeffrey A, and Zhong Wang. 2011. “Next-Generation Transcriptome Assembly.” *Nature Reviews. Genetics* 12 (10): 671–82. https://doi.org/10.1038/nrg3068.

Martínez-Jiménez, Francisco, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, et al. 2020. “A Compendium of Mutational Cancer Driver Genes.” *Nature Reviews Cancer*. Nature Research. https://doi.org/10.1038/s41568-020-0290-x.

Mastrangelo, Giuseppe, Emanuela Fadda, and Vita Marzia. 1996. “Polycyclic Aromatic Hydrocarbons and Cancer in Man.” *Environmental Health Perspectives*. Public Health Services, US Dept of Health and Human Services. https://doi.org/10.1289/ehp.961041166.

Mauvais-Jarvis, Franck, Noel Bairey Merz, Peter J. Barnes, Roberta D. Brinton, Juan Jesus Carrero, Dawn L. DeMeo, Geert J. De Vries, et al. 2020. “Sex and Gender: Modifiers of Health, Disease, and Medicine.” *The Lancet*. Lancet Publishing Group. https://doi.org/10.1016/S0140-6736(20)31561-0.

McCartney, Gerry, Lamia Mahmood, Alastair H. Leyland, G. David Batty, and Kate Hunt. 2011. “Contribution of Smoking-Related and Alcohol-Related Deaths to the Gender Gap in Mortality: Evidence from 30 European Countries.” *Tobacco Control* 20 (2): 166–68. https://doi.org/10.1136/tc.2010.037929.

McDermott, Jason E., Osama A. Arshad, Vladislav A. Petyuk, Yi Fu, Marina A. Gritsenko, Therese R. Clauss, Ronald J. Moore, et al. 2020. “Proteogenomic Characterization of Ovarian HGSC Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal Instability.” *Cell Reports Medicine* 1 (1): 100004. https://doi.org/10.1016/j.xcrm.2020.100004.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R.S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant

Effect Predictor." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0974-4.

McShane, Erik, Celine Sin, Henrik Zauber, Jonathan N. Wells, Neysan Donnelly, Xi Wang, Jingyi Hou, et al. 2016. "Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation." *Cell* 167 (3): 803-815.e21. https://doi.org/10.1016/j.cell.2016.09.015.

Medicine, Institute of. 2012. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Edited by Christine M Micheel, Sharly J Nass, and Gilbert S Omenn. Washington, DC: The National Academies Press. https://doi.org/10.17226/13297.

Meindl, Alfons, Heide Hellebrand, Constanze Wiek, Verena Erven, Barbara Wappenschmidt, Dieter Niederacher, Marcel Freund, et al. 2010. "Germline Mutations in Breast and Ovarian Cancer Pedigrees Establish RAD51C as a Human Cancer Susceptibility Gene." *Nature Genetics* 42 (5): 410–14. https://doi.org/10.1038/ng.569.

Melé, Marta, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, et al. 2015. "The Human Transcriptome across Tissues and Individuals." *Science* 348 (6235): 660–65. https://doi.org/10.1126/science.aaa0355.

Melo, Sonia A., Catia Moutinho, Santiago Ropero, George A. Calin, Simona Rossi, Riccardo Spizzo, Agustin F. Fernandez, et al. 2010. "A Genetic Defect in Exportin-5 Traps Precursor MicroRNAs in the Nucleus of Cancer Cells." *Cancer Cell* 18 (4): 303–15. https://doi.org/10.1016/j.ccr.2010.09.007.

Mermel, Craig H., Steven E. Schumacher, Barbara Hill, Matthew L. Meyerson, Rameen Beroukhim, and Gad Getz. 2011. "GISTIC2.0 Facilitates Sensitive and Confident Localization of the Targets of Focal Somatic Copy-Number Alteration in Human Cancers." *Genome Biology* 12 (4). https://doi.org/10.1186/gb-2011-12-4-r41.

Mertins, Philipp, D. R. Mani, Kelly V. Ruggles, Michael A. Gillette, Karl R. Clauser, Pei Wang, Xianlong Wang, et al. 2016. "Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer." *Nature* 534 (7605): 55–62. https://doi.org/10.1038/nature18003.

Messmer, Tobias, Ferdinand von Meyenn, Aurora Savino, Fátima Santos, Hisham Mohammed, Aaron Tin Long Lun, John C. Marioni, and Wolf Reik. 2019. "Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution." *Cell Reports* 26 (4): 815-824.e4. https://doi.org/10.1016/j.celrep.2018.12.099.

Meulen, Joni Van Der, Viraj Sanghvi, Konstantinos Mavrakis, Kaat Durinck, Fang Fang, Filip Matthijssens, Pieter Rondou, et al. 2015. "The H3K27me3 Demethylase UTX Is a Gender-Specific Tumor Suppressor in T-Cell Acute Lymphoblastic Leukemia." *Blood* 125 (1): 13–21. https://doi.org/10.1182/blood-2014-05-577270.

Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E.

Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. “Computational Correction of Copy Number Effect Improves Specificity of CRISPR-Cas9 Essentiality Screens in Cancer Cells.” *Nature Genetics* 49 (12): 1779–84. https://doi.org/10.1038/ng.3984.

Miller, Chad J., and Benjamin E. Turk. 2018. “Homing in: Mechanisms of Substrate Targeting by Protein Kinases.” *Trends in Biochemical Sciences*. Elsevier Ltd. https://doi.org/10.1016/j.tibs.2018.02.009.

Mirauta, Bogdan Andrei, Daniel D. Seaton, Dalila Bensaddek, Alejandro Brenes, Marc Jan Bonder, Helena Kilpinen, Chukwuma A. Agu, et al. 2020. “Population-Scale Proteome Variation in Human Induced Pluripotent Stem Cells.” *ELife* 9 (August): 1–22. https://doi.org/10.7554/ELIFE.57390.

Mishra, Gopa R., M. Suresh, K. Kumaran, N. Kannabiran, Shubha Suresh, P. Bala, K. Shivakumar, et al. 2006. “Human Protein Reference Database--2006 Update.” *Nucleic Acids Research* 34 (Database issue). https://doi.org/10.1093/nar/gkj141.

Mittelstrass, Kirstin, Janina S. Ried, Zhonghao Yu, Jan Krumsiek, Christian Gieger, Cornelia Prehn, Werner Roemisch-Margl, et al. 2011. “Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers.” *PLoS Genetics* 7 (8). https://doi.org/10.1371/journal.pgen.1002215.

Mittendorfer, Bettina. 2005. “Sexual Dimorphism in Human Lipid Metabolism.” *Journal of Nutrition*. American Institute of Nutrition. https://doi.org/10.1093/jn/135.4.681.

Mitternacht, Simon. 2016. “FreeSASA: An Open Source C Library for Solvent Accessible Surface Area Calculations.” *F1000Research* 5. https://doi.org/10.12688/f1000research.7931.1.

Mok, Janine, Xiaowei Zhu, and Michael Snyder. 2011. “Dissecting Phosphorylation Networks: Lessons Learned from Yeast.” *Expert Review of Proteomics*. Expert Rev Proteomics. https://doi.org/10.1586/epr.11.64.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Mosca, Roberto, Arnaud Céol, and Patrick Aloy. 2013. “Interactome3D: Adding Structural Details to Protein Networks.” *Nature Methods* 10 (1): 47–53. https://doi.org/10.1038/nmeth.2289.

Mousavi, Mohammad Javad, Mahdi Mahmoudi, and Somayeh Ghotloo. 2020. “Escape from X Chromosome Inactivation and Female Bias of Autoimmune Diseases.” *Molecular Medicine*. BioMed Central Ltd. https://doi.org/10.1186/s10020-020-00256-1.

Muir, Brian, and Leonard Nunney. 2015. “The Expression of Tumour Suppressors and Proto-Oncogenes in Tissues Susceptible to Their Hereditary Cancers.” *British Journal of Cancer* 113 (2): 345–53. https://doi.org/10.1038/bjc.2015.205.

Mun, Dong Gi, Jinhyuk Bhin, Sangok Kim, Hyunwoo Kim, Jae Hun Jung, Yeonjoo Jung, Ye Eun Jang, et al. 2019. “Proteogenomic Characterization of Human Early-Onset Gastric Cancer.” *Cancer Cell* 35 (1): 111-124.e10. https://doi.org/10.1016/j.ccell.2018.12.003.

Mushegian, A. R., and E. V. Koonin. 1996. “Gene Order Is Not Conserved in Bacterial Evolution.” *Trends in Genetics : TIG*. Trends Genet. https://doi.org/10.1016/0168-9525(96)20006-X.

Muzny, Donna M., Matthew N. Bainbridge, Kyle Chang, Huyen H. Dinh, Jennifer A. Drummond, Gerald Fowler, Christie L. Kovar, et al. 2012. “Comprehensive Molecular Characterization of Human Colon and Rectal Cancer.” *Nature* 487 (7407): 330–37. https://doi.org/10.1038/nature11252.

Naqvi, Sahin, Alexander K. Godfrey, Jennifer F. Hughes, Mary L. Goodheart, Richard N. Mitchell, and David C. Page. 2019. “Conservation, Acquisition, and Functional Impact of Sex-Biased Gene Expression in Mammals.” *Science* 365 (6450). https://doi.org/10.1126/science.aaw7317.

Naugler, Willscott E., Toshiharu Sakurai, Sunhwa Kim, Shin Maeda, Kyoung Hyun Kim, Ahmed M. Elsharkawy, and Michael Karin. 2007. “Gender Disparity in Liver Cancer Due to Sex Differences in MyD88-Dependent IL-6 Production.” *Science* 317 (5834): 121–24. https://doi.org/10.1126/science.1140485.

Negrini, Simona, Vassilis G. Gorgoulis, and Thanos D. Halazonetis. 2010. “Genomic Instability an Evolving Hallmark of Cancer.” *Nature Reviews Molecular Cell Biology*. Nat Rev Mol Cell Biol. https://doi.org/10.1038/nrm2858.

Nesvizhskii, Alexey I. 2014. “Proteogenomics: Concepts, Applications and Computational Strategies.” *Nature Methods*. Nature Publishing Group. https://doi.org/10.1038/NMETH.3144.

Newman, Robert H., Jin Zhang, and Heng Zhu. 2014. “Toward a Systems-Level View of Dynamic Phosphorylation Networks.” *Frontiers in Genetics*. Frontiers Research Foundation. https://doi.org/10.3389/fgene.2014.00263.

Ngo, S. T., F. J. Steyn, and P. A. McCombe. 2014. “Gender Differences in Autoimmune Disease.” *Frontiers in Neuroendocrinology*. Academic Press Inc. https://doi.org/10.1016/j.yfrne.2014.04.004.

Nishi, Hafumi, Kosuke Hashimoto, and Anna R. Panchenko. 2011. “Phosphorylation in Protein-Protein Binding: Effect on Stability and Function.” *Structure* 19 (12): 1807–15. https://doi.org/10.1016/j.str.2011.09.021.

Nishi, Hafumi, Alexey Shaytan, and Anna R. Panchenko. 2014. “Physicochemical Mechanisms of Protein Regulation by Phosphorylation.” *Frontiers in Genetics*. Frontiers Research Foundation. https://doi.org/10.3389/fgene.2014.00270.

Nowell, Peter C. 1976. “The Clonal Evolution of Tumor Cell Populations.” *Science* 194

(4260): 23–28. https://doi.org/10.1126/science.959840.

Nusinow, David P., John Szpyt, Mahmoud Ghandi, Christopher M. Rose, E. Robert McDonald, Marian Kalocsay, Judit Jané-Valbuena, et al. 2020. “Quantitative Proteomics of the Cancer Cell Line Encyclopedia.” *Cell* 180 (2): 387-402.e16. https://doi.org/10.1016/j.cell.2019.12.023.

Ober, Carole, Dagan A Loisel, and Yoav Gilad. 2008. “Sex-Specific Genetic Architecture of Human Disease.” *Nature Reviews. Genetics* 9 (12): 911–22. https://doi.org/10.1038/nrg2415.

Ochoa, David, Andrew F. Jarnuczak, Cristina Viéitez, Maja Gehre, Margaret Soucheray, André Mateus, Askar A. Kleefeldt, et al. 2020. “The Functional Landscape of the Human Phosphoproteome.” *Nature Biotechnology* 38 (3): 365–73. https://doi.org/10.1038/s41587-019-0344-3.

Ochoa, David, Mindaugas Jonikas, Robert T Lawrence, Bachir El Debs, Joel Selkrig, Athanasios Typas, Judit Villén, Silvia DM Santos, and Pedro Beltrao. 2016. “An Atlas of Human Kinase Regulation.” *Molecular Systems Biology* 12 (12): 888. https://doi.org/10.15252/msb.20167295.

Ohh, Michael, Cheol Won Park, Mircea Ivan, Michael A. Hoffman, Tae You Kim, L. Eric Huang, Nikola Pavletich, Vincent Chau, and William G. Kaelin. 2000. “Ubiquitination of Hypoxia-Inducible Factor Requires Direct Binding to the β-Domain of the von Hippel - Lindau Protein.” *Nature Cell Biology* 2 (7): 423–27. https://doi.org/10.1038/35017054.

Oliner, J. D., K. W. Kinzler, P. S. Meltzer, D. L. George, and B. Vogelstein. 1992. “Amplification of a Gene Encoding a P53-Associated Protein in Human Sarcomas.” *Nature* 358 (6381): 80–83. https://doi.org/10.1038/358080a0.

Olsen, Jesper V., Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. 2006. “Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks.” *Cell* 127 (3): 635–48. https://doi.org/10.1016/j.cell.2006.09.026.

Olzscha, Heidi, Sonya M. Schermann, Andreas C. Woerner, Stefan Pinkert, Michael H. Hecht, Gian G. Tartaglia, Michele Vendruscolo, Manajit Hayer-Hartl, F. Ulrich Hartl, and R. Martin Vabulas. 2011. “Amyloid-like Aggregates Sequester Numerous Metastable Proteins with Essential Cellular Functions.” *Cell* 144 (1): 67–78. https://doi.org/10.1016/j.cell.2010.11.050.

Ong, Shao En, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. 2002. “Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.” *Molecular & Cellular Proteomics : MCP* 1 (5): 376–86. https://doi.org/10.1074/mcp.M200025-MCP200.

Ong, Shao En, and Matthias Mann. 2005. “Mass Spectrometry–Based Proteomics Turns Quantitative.” *Nature Chemical Biology* 1 (5): 252–62. https://doi.org/10.1038/nchembio736.

Ongen, Halit, Claus L. Andersen, Jesper B. Bramsen, Bodil Oster, Mads H. Rasmussen, Pedro G. Ferreira, Juan Sandoval, et al. 2014. “Putative Cis-Regulatory Drivers in Colorectal Cancer.” *Nature* 512 (1): 87–90. https://doi.org/10.1038/nature13602.

Ortsäter, Henrik, Nina Grankvist, Richard E. Honkanen, and Åke Sjöholm. 2014. “Protein Phosphatases in Pancreatic Islets.” *Journal of Endocrinology*. BioScientifica Ltd. https://doi.org/10.1530/JOE-14-0002.

Osmanbeyoglu, Hatice U., Eneda Toska, Carmen Chan, José Baselga, and Christina S. Leslie. 2017. “Pancancer Modelling Predicts the Context-Specific Impact of Somatic Mutations on Transcriptional Programs.” *Nature Communications* 8 (January). https://doi.org/10.1038/ncomms14249.

Owens, Marilyn A., Bruce C. Horten, and Moacyr M. Da Silva. 2004. “HER2 Amplification Ratios by Fluorescence in Situ Hybridization and Correlation with Immunohistochemistry in a Cohort of 6556 Breast Cancer Tissues.” *Clinical Breast Cancer* 5 (1): 63–69. https://doi.org/10.3816/CBC.2004.n.011.

Paschal, Catherine Randall, John Maciejowski, and Prasad V. Jallepalli. 2012. “A Stringent Requirement for PlK1 T210 Phosphorylation during K-Fiber Assembly and Chromosome Congression.” *Chromosoma* 121 (6): 565–72. https://doi.org/10.1007/s00412-012-0375-8.

Passarelli, Michael N., Amanda I. Phipps, John D. Potter, Karen W. Makar, Anna E. Coghill, Karen J. Wernli, Emily White, et al. 2013. “Common Single-Nucleotide Polymorphisms in the Estrogen Receptor β Promoter Are Associated with Colorectal Cancer Survival in Postmenopausal Women.” *Cancer Research* 73 (2): 767–75. https://doi.org/10.1158/0008-5472.CAN-12-2484.

Paulovich, Amanda G., and Jeffrey R. Whiteaker. 2016. “Quantifying the Human Proteome.” *Nature Biotechnology*. Nature Publishing Group. https://doi.org/10.1038/nbt.3695.

Pavelka, Norman, Giulia Rancati, Jin Zhu, William D. Bradford, Anita Saraf, Laurence Florens, Brian W. Sanderson, Gaye L. Hattem, and Rong Li. 2010. “Aneuploidy Confers Quantitative Proteome Changes and Phenotypic Variation in Budding Yeast.” *Nature* 468 (7321): 321–25. https://doi.org/10.1038/nature09529.

Peles, E., R. Lamprecht, R. Ben-Levy, E. Tzahar, and Y. Yarden. 1992. “Regulated Coupling of the Neu Receptor to Phosphatidylinositol 3’-Kinase and Its Release by Oncogenic Activation.” *Journal of Biological Chemistry* 267 (17): 12266–74. https://doi.org/10.1016/s0021-9258(19)49834-7.

Perfetto, Livia, Leonardo Briganti, Alberto Calderone, Andrea Cerquone Perpetuini, Marta

Iannuccelli, Francesca Langone, Luana Licata, et al. 2016. “SIGNOR: A Database of Causal Relationships between Biological Entities.” *Nucleic Acids Research* 44 (D1): D548–54. https://doi.org/10.1093/nar/gkv1048.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. “StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads.” *Nature Biotechnology* 33 (3): 290–95. https://doi.org/10.1038/nbt.3122.

Petralia, Francesca, Nicole Tignor, Boris Reva, Mateusz Koptyra, Shrabanti Chowdhury, Dmitry Rykunov, Azra Krek, et al. 2020. “Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer.” *Cell* 183 (7): 1962-1985.e31. https://doi.org/10.1016/j.cell.2020.10.044.

Petryszak, Robert, Tony Burdett, Benedetto Fiorelli, Nuno A. Fonseca, Mar Gonzalez-Porta, Emma Hastings, Wolfgang Huber, et al. 2014. “Expression Atlas Update - A Database of Gene and Transcript Expression from Microarray- and Sequencing-Based Functional Genomics Experiments.” *Nucleic Acids Research* 42 (D1). https://doi.org/10.1093/nar/gkt1270.

Pinkse, Martijn W.H., Pauliina M. Uitto, Martijn J. Hilhorst, Bert Ooms, and Albert J.R. Heck. 2004. “Selective Isolation at the Femtomole Level of Phosphopeptides from Proteolytic Digests Using 2D-NanoLC-ESI-MS/MS and Titanium Oxide Precolumns.” *Analytical Chemistry* 76 (14): 3935–43. https://doi.org/10.1021/ac0498617.

Pleasance, Erin D., R. Keira Cheetham, Philip J. Stephens, David J. McBride, Sean J. Humphray, Chris D. Greenman, Ignacio Varela, et al. 2010. “A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome.” *Nature* 463 (7278): 191–96. https://doi.org/10.1038/nature08658.

Posewitz, Matthew C., and Paul Tempst. 1999. “Immobilized Gallium(III) Affinity Chromatography of Phosphopeptides.” *Analytical Chemistry* 71 (14): 2883–92. https://doi.org/10.1021/ac981409y.

Pozniak, Yair, Nora Balint-Lahat, Jan Daniel Rudolph, Cecilia Lindskog, Rotem Katzir, Camilla Avivi, Fredrik Pontén, Eytan Ruppin, Iris Barshack, and Tamar Geiger. 2016. “System-Wide Clinical Proteomics of Breast Cancer Reveals Global Remodeling of Tissue Homeostasis.” *Cell Systems* 2 (3): 172–84. https://doi.org/10.1016/j.cels.2016.02.001.

Pranavathiyani, G., Raja Rajeswary Thanmalagan, Naorem Leimarembi Devi, and Amouda Venkatesan. 2019. “Integrated Transcriptome Interactome Study of Oncogenes and Tumor Suppressor Genes in Breast Cancer.” *Genes and Diseases* 6 (1): 78–87. https://doi.org/10.1016/j.gendis.2018.10.004.

Prus, Gabriela, Annabelle Hoegl, Brian T. Weinert, and Chunaram Choudhary. 2019.

"Analysis and Interpretation of Protein Post-Translational Modification Site Stoichiometry." *Trends in Biochemical Sciences*. Elsevier Ltd. https://doi.org/10.1016/j.tibs.2019.06.003.

Rahbari, Reza, Lisa Zhang, and Electron Kebebew. 2010. "Thyroid Cancer Gender Disparity." *Future Oncology*. Future Oncol. https://doi.org/10.2217/fon.10.127.

Rahman, Mumtahena, Laurie K. Jackson, W. Evan Johnson, Dean Y. Li, Andrea H. Bild, and Stephen R. Piccolo. 2015. "Alternative Preprocessing of RNA-Sequencing Data in the Cancer Genome Atlas Leads to Improved Analysis Results." *Bioinformatics* 31 (22): 3666–72. https://doi.org/10.1093/bioinformatics/btv377.

Raman, Priyadarshini, and Ronald J. Koenig. 2014. "Pax-8-PPAR-γ 3 Fusion Protein in Thyroid Carcinoma." *Nature Reviews Endocrinology*. Nature Publishing Group. https://doi.org/10.1038/nrendo.2014.115.

Ramos, Alex H., Lee Lichtenstein, Manaswi Gupta, Michael S. Lawrence, Trevor J. Pugh, Gordon Saksena, Matthew Meyerson, and Gad Getz. 2015. "Oncotator: Cancer Variant Annotation Tool." *Human Mutation* 36 (4): E2423–29. https://doi.org/10.1002/humu.22771.

Rawlik, Konrad, Oriol Canela-Xandri, and Albert Tenesa. 2016. "Evidence for Sex-Specific Genetic Architectures across a Spectrum of Human Complex Traits." *Genome Biology* 17 (1): 1–8. https://doi.org/10.1186/s13059-016-1025-x.

Reimand, Jüri, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, et al. 2019. "Pathway Enrichment Analysis and Visualization of Omics Data Using g:Profiler, GSEA, Cytoscape and EnrichmentMap." *Nature Protocols* 14 (2): 482–517. https://doi.org/10.1038/s41596-018-0103-9.

Reinders, Joerg, and Albert Sickmann. 2005. "State-of-the-Art in Phosphoproteomics." *Proteomics*. Proteomics. https://doi.org/10.1002/pmic.200401289.

Reizel, Yitzhak, Adam Spiro, Ofra Sabag, Yael Skversky, Merav Hecht, Ilana Keshet, Benjamin P. Berman, and Howard Cedar. 2015. "Gender-Specific Postnatal Demethylation and Establishment of Epigenetic Memory." *Genes and Development* 29 (9): 923–33. https://doi.org/10.1101/gad.259309.115.

Rikova, Klarisa, Ailan Guo, Qingfu Zeng, Anthony Possemato, Jian Yu, Herbert Haack, Julie Nardone, et al. 2007. "Global Survey of Phosphotyrosine Signaling Identifies Oncogenic Kinases in Lung Cancer." *Cell* 131 (6): 1190–1203. https://doi.org/10.1016/j.cell.2007.11.025.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Rivlin, Noa, Ran Brosh, Moshe Oren, and Varda Rotter. 2011. “Mutations in the P53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis.” *Genes and Cancer*. Genes Cancer. https://doi.org/10.1177/1947601911408889.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11: R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Rodriguez, Henry, Jean Claude Zenklusen, Louis M. Staudt, James H. Doroshow, and Douglas R. Lowy. 2021. “The next Horizon in Precision Oncology: Proteogenomics to Inform Cancer Diagnosis and Treatment.” *Cell*. Elsevier B.V. https://doi.org/10.1016/j.cell.2021.02.055.

Ross, Philip L., Yulin N. Huang, Jason N. Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, et al. 2004. “Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents.” *Molecular and Cellular Proteomics* 3 (12): 1154–69. https://doi.org/10.1074/mcp.M400129-MCP200.

Roumeliotis, Theodoros I., Steven P. Williams, Emanuel Gonçalves, Clara Alsinet, Martin Del Castillo Velasco-Herrera, Nanne Aben, Fatemeh Zamanzad Ghavidel, et al. 2017. “Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells.” *Cell Reports* 20 (9): 2201–14. https://doi.org/10.1016/j.celrep.2017.08.010.

Rush, John, Albrecht Moritz, Kimberly A. Lee, Ailan Guo, Valerie L. Goss, Erik J. Spek, Hui Zhang, Xiang Ming Zha, Roberto D. Polakiewicz, and Michael J. Comb. 2005. “Immunoaffinity Profiling of Tyrosine Phosphorylation in Cancer Cells.” *Nature Biotechnology* 23 (1): 94–101. https://doi.org/10.1038/nbt1046.

Ruttenberg, Brian E., Trairak Pisitkun, Mark A. Knepper, and Jason D. Hoffert. 2008. “PhosphoScore: An Open-Source Phosphorylation Site Assignment Too! For MSn Data.” *Journal of Proteome Research* 7 (7): 3054–59. https://doi.org/10.1021/pr800169k.

Ryan, Colm J., Susan Kennedy, Ilirjana Bajrami, David Matallanas, and Christopher J. Lord. 2017. “A Compendium of Co-Regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events.” *Cell Systems* 5 (4): 399-409.e5. https://doi.org/10.1016/j.cels.2017.09.011.

Saad, Everardo D., Xavier Paoletti, Tomasz Burzykowski, and Marc Buyse. 2017. “Precision Medicine Needs Randomized Clinical Trials.” *Nature Reviews Clinical Oncology*. Nature Publishing Group. https://doi.org/10.1038/nrclinonc.2017.8.

Saha, Ashis, Yungil Kim, Ariel D.H. Gewirtz, Brian Jo, Chuan Gao, Ian C. McDowell, Barbara E. Engelhardt, et al. 2017. “Co-Expression Networks Reveal the Tissue-Specific Regulation of Transcription and Splicing.” *Genome Research* 27 (11): 1843–58. https://doi.org/10.1101/gr.216721.116.

Sancar, Aziz, Laura A. Lindsey-Boltz, Keziban Ünsal-Kaçmaz, and Stuart Linn. 2004. “Molecular Mechanisms of Mammalian DNA Repair and the DNA Damage Checkpoints.” *Annual Review of Biochemistry*. Annu Rev Biochem. https://doi.org/10.1146/annurev.biochem.73.011303.073723.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67. https://doi.org/10.1073/pnas.74.12.5463.

Saul, R. L., and B. N. Ames. 1986. “Background Levels of DNA Damage in the Population.” *Basic Life Sciences* 38: 529–35. https://doi.org/10.1007/978-1-4615-9462-8_55.

Savitski, Mikhail M., Michael L. Nielsen, and Roman A. Zubarev. 2006. “ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-Translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures.” *Molecular and Cellular Proteomics* 5 (5): 935–48. https://doi.org/10.1074/mcp.T500034-MCP200.

Scarselli, Alberto, Marisa Corfiati, Davide Di Marzio, Alessandro Marinaccio, and Sergio Iavicoli. 2018. “Gender Differences in Occupational Exposure to Carcinogens among Italian Workers.” *BMC Public Health* 18 (1). https://doi.org/10.1186/s12889-018-5332-x.

Schaefer, Carl F., Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. 2009. “PID: The Pathway Interaction Database.” *Nucleic Acids Research* 37 (SUPPL. 1). https://doi.org/10.1093/nar/gkn653.

Schimek, Michael G., Eva Budinská, Karl G. Kugler, Vendula Švendová, Jie Ding, and Shili Lin. 2015. “TopKLists: A Comprehensive R Package for Statistical Inference, Stochastic Aggregation, and Visualization of Multiple Omics Ranked Lists.” *Statistical Applications in Genetics and Molecular Biology* 14 (3): 311–16. https://doi.org/10.1515/sagmb-2014-0093.

Schoenberg, Daniel R., and Lynne E. Maquat. 2012. “Regulation of Cytoplasmic MRNA Decay.” *Nature Reviews Genetics*. Nat Rev Genet. https://doi.org/10.1038/nrg3160.

Schwämmle, Veit, Ileana Rodríguez León, and Ole Nørregaard Jensen. 2013. “Assessment and Improvement of Statistical Tools for Comparative Proteomics Analysis of Sparse Data Sets with Few Experimental Replicates.” *Journal of Proteome Research* 12 (9): 3874–83. https://doi.org/10.1021/pr400045u.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. “Mclust 5:

Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *R Journal* 8 (1): 289–317. https://doi.org/10.32614/rj-2016-021.

Seyfried, Nicholas T., Eric B. Dammer, Vivek Swarup, Divya Nandakumar, Duc M. Duong, Luming Yin, Qiudong Deng, et al. 2017. "A Multi-Network Approach Identifies Protein-Specific Co-Expression in Asymptomatic and Symptomatic Alzheimer's Disease." *Cell Systems* 4 (1): 60-72.e4. https://doi.org/10.1016/j.cels.2016.11.006.

Sharma, Anchal, Chuan Jiang, and Subhajyoti De. 2018. "Dissecting the Sources of Gene Expression Variation in a Pan-Cancer Analysis Identifies Novel Regulatory Mutations." *Nucleic Acids Research* 46 (9): 4370–81. https://doi.org/10.1093/nar/gky271.

Sharp, Andrew J., Elisavet Stathaki, Eugenia Migliavacca, Manisha Brahmachary, Stephen B. Montgomery, Yann Dupre, and Stylianos E. Antonarakis. 2011. "DNA Methylation Profiles of Human Active and Inactive X Chromosomes." *Genome Research* 21 (10): 1592–1600. https://doi.org/10.1101/gr.112680.110.

Shen, Ying, Xin Li, Dandan Dong, Bin Zhang, Yanru Xue, and Peng Shang. 2018. "Transferrin Receptor 1 in Cancer: A New Sight for Cancer Therapy." *American Journal of Cancer Research* 8 (6): 916–31. http://www.ncbi.nlm.nih.gov/pubmed/30034931.

Shendure, Jay, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. 2017. "DNA Sequencing at 40: Past, Present and Future." *Nature* 550 (7676). https://doi.org/10.1038/nature24286.

Sherman, Rachel M., and Steven L. Salzberg. 2020. "Pan-Genomics in the Human Genome Era." *Nature Reviews Genetics*. Nature Research. https://doi.org/10.1038/s41576-020-0210-7.

Shimizu, I., M. Yasuda, Y. Mizobuchi, Y. R. Ma, F. Liu, M. Shiba, T. Horie, and S. Ito. 1998. "Suppressive Effect of Oestradiol on Chemical Hepatocarcinogenesis in Rats." *Gut* 42 (1): 112–19. https://doi.org/10.1136/gut.42.1.112.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. 2018. "Cancer Statistics, 2018." *CA: A Cancer Journal for Clinicians* 68 (1): 7–30. https://doi.org/10.3322/caac.21442.

Sinitcyn, Pavel, Jan Daniel Rudolph, and Jürgen Cox. 2018. "Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data." *Annual Review of Biomedical Data Science* 1 (1): 207–34. https://doi.org/10.1146/annurev-biodatasci-080917-013516.

Slamon, Dennis J., William Godolphin, Lovell A. Jones, John A. Holt, Steven G. Wong, Duane E. Keith, Wendy J. Levin, et al. 1989. "Studies of the HER-2/Neu Proto-Oncogene in Human Breast and Ovarian Cancer." *Science* 244 (4905): 707–12. https://doi.org/10.1126/science.2470152.

Slotta, Douglas J., Melinda A. McFarland, and Sanford P. Markey. 2010. "MassSieve: Panning MS/MS Peptide Data for Proteins." *Proteomics* 10 (16): 3035–39.

https://doi.org/10.1002/pmic.200900370.

Smith, Lloyd M., and Neil L. Kelleher. 2013. “Proteoform: A Single Term Describing Protein Complexity.” *Nature Methods*. Nat Methods. https://doi.org/10.1038/nmeth.2369.

Smits, Alexander J.J., J. Alain Kummer, Peter C. De Bruin, Mijke Bol, Jan G. Van Den Tweel, Kees A. Seldenrijk, Stefan M. Willems, et al. 2014. “The Estimation of Tumor Cell Percentage for Molecular Testing by Pathologists Is Not Accurate.” *Modern Pathology* 27 (2): 168–74. https://doi.org/10.1038/modpathol.2013.134.

Somyajit, Kumar, Shreelakshmi Subramanya, and Ganesh Nagaraju. 2010. “RAD51C: A Novel Cancer Susceptibility Gene Is Linked to Fanconi Anemia and Breast Cancer.” *Carcinogenesis* 31 (12): 2031–38. https://doi.org/10.1093/carcin/bgq210.

Sondka, Zbyslaw, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. 2018. “The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers.” *Nature Reviews Cancer*. Nature Publishing Group. https://doi.org/10.1038/s41568-018-0060-1.

Soneson, Charlotte, and Mauro Delorenzi. 2013. “A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data.” *BMC Bioinformatics* 14: 91. https://doi.org/10.1186/1471-2105-14-91.

Soto, Claudio, and Sandra Pritzkow. 2018. “Protein Misfolding, Aggregation, and Conformational Strains in Neurodegenerative Diseases.” *Nature Neuroscience*. Nature Publishing Group. https://doi.org/10.1038/s41593-018-0235-9.

Sousa, Abel, Emanuel Gonçalves, Bogdan Mirauta, David Ochoa, Oliver Stegle, and Pedro Beltrao. 2019. “Multi-Omics Characterization of Interactionmediated Control of Human Protein Abundance Levels.” *Molecular and Cellular Proteomics* 18 (8): S114–25. https://doi.org/10.1074/mcp.RA118.001280.

Spring, Laura M., Mark L. Zangardi, Beverly Moy, and Aditya Bardia. 2017. “Clinical Management of Potential Toxicities and Drug Interactions Related to Cyclin-Dependent Kinase 4/6 Inhibitors in Breast Cancer: Practical Considerations and Recommendations.” *The Oncologist* 22 (9): 1039–48. https://doi.org/10.1634/theoncologist.2017-0142.

Stanley, J. A., M. M. Aruldhas, M. Chandrasekaran, R. Neelamohan, E. Suthagar, K. Annapoorna, S. Sharmila, et al. 2012. “Androgen Receptor Expression in Human Thyroid Cancer Tissues: A Potential Mechanism Underlying the Gender Bias in the Incidence of Thyroid Cancers.” *Journal of Steroid Biochemistry and Molecular Biology* 130 (1–2): 105–24. https://doi.org/10.1016/j.jsbmb.2012.02.004.

Stark, Chris, Bobby Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. “BioGRID: A General Repository for Interaction Datasets.” *Nucleic Acids Research* 34 (Database issue). https://doi.org/10.1093/nar/gkj109.

Stark, Rory, Marta Grzelak, and James Hadfield. 2019. “RNA Sequencing: The Teenage Years.” *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/s41576-019-0150-2.

Stingele, Silvia, Gabriele Stoehr, Karolina Peplowska, Jürgen Cox, Matthias Mann, and Zuzana Storchova. 2012. “Global Analysis of Genome, Transcriptome and Proteome Reveals the Response to Aneuploidy in Human Cells.” *Molecular Systems Biology* 8 (1): 608. https://doi.org/10.1038/msb.2012.40.

Stransky, Nicolas, Mahmoud Ghandi, Gregory V. Kryukov, Levi A. Garraway, Joseph Lehár, Manway Liu, Dmitriy Sonkin, et al. 2015. “Pharmacogenomic Agreement between Two Cancer Cell Line Data Sets.” *Nature* 528 (7580): 84–87. https://doi.org/10.1038/nature15736.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. “The Cancer Genome.” *Nature*. Nature. https://doi.org/10.1038/nature07943.

Su, Zhengming, Duqun Chen, Enpu Zhang, Yifan Li, Zuhu Yu, Min Shi, Zhimao Jiang, et al. 2015. “MicroRNA-509-3p Inhibits Cancer Cell Proliferation and Migration by Targeting the Mitogen-Activated Protein Kinase Kinase Kinase 8 Oncogene in Renal Cell Carcinoma.” *Molecular Medicine Reports* 12 (1): 1535–43. https://doi.org/10.3892/mmr.2015.3498.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Sugiyama, Naoyuki, Haruna Imamura, and Yasushi Ishihama. 2019. “Large-Scale Discovery of Substrates of the Human Kinome.” *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-46385-4.

Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.” *CA: A Cancer Journal for Clinicians* 71 (3): 209–49. https://doi.org/10.3322/caac.21660.

Sutandy, F. X.Reymond, Jiang Qian, Chien Sheng Chen, and Heng Zhu. 2013. “Overview of Protein Microarrays.” *Current Protocols in Protein Science* Chapter 27 (SUPPL.72). https://doi.org/10.1002/0471140864.ps2701s72.

Taguchi, Keiko, and Masayuki Yamamoto. 2017. “The KEAP1NRF2 System in Cancer.” *Frontiers in Oncology*. Frontiers Media S.A. https://doi.org/10.3389/fonc.2017.00085.

Tang, Yun Chi, and Angelika Amon. 2013. “Gene Copy-Number Alterations: A Cost-Benefit

Analysis." *Cell*. Elsevier B.V. https://doi.org/10.1016/j.cell.2012.11.043.

Tarasov, Artem, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. 2015. "Sambamba: Fast Processing of NGS Alignment Formats." *Bioinformatics* 31 (12): 2032–34. https://doi.org/10.1093/bioinformatics/btv098.

Tarazona, Sonia, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. 2011. "Differential Expression in RNA-Seq: A Matter of Depth." *Genome Research* 21 (12): 2213–23. https://doi.org/10.1101/gr.124321.111.

Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. "COSMIC: The Catalogue Of Somatic Mutations In Cancer." *Nucleic Acids Research* 47 (D1): D941–47. https://doi.org/10.1093/nar/gky1015.

Telonis, Aristeidis G., Phillipe Loher, Rogan Magee, Venetia Pliatsika, Eric Londin, Yohei Kirino, and Isidore Rigoutsos. 2019. "TRNA Fragments Show Intertwining with MRNAs of Specific Repeat Content and Have Links to Disparities." *Cancer Research* 79 (12): 3034–49. https://doi.org/10.1158/0008-5472.CAN-19-0789.

Tevfik Dorak, M., and Ebru Karpuzoglu. 2012. "Gender Differences in Cancer Susceptibility: An Inadequately Addressed Issue." *Frontiers in Genetics* 3 (NOV): 1–11. https://doi.org/10.3389/fgene.2012.00268.

Thompson, Andrew, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. 2003. "Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS." *Analytical Chemistry* 75 (8): 1895–1904. https://doi.org/10.1021/ac0262560.

Thompson, Andrew, Nikolai Wölmer, Sasa Koncarevic, Stefan Selzer, Gitte Böhm, Harald Legner, Peter Schmid, et al. 2019. "TMTpro: Design, Synthesis, and Initial Evaluation of a Proline-Based Isobaric 16-Plex Tandem Mass Tag Reagent Set." *Analytical Chemistry* 91 (24): 15941–50. https://doi.org/10.1021/acs.analchem.9b04474.

Thorsson, Vésteinn, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai Hsien Ou Yang, Eduard Porta-Pardo, et al. 2018. "The Immune Landscape of Cancer." *Immunity* 48 (4): 812-830.e14. https://doi.org/10.1016/j.immuni.2018.03.023.

Tomasetti, Cristian, and Bert Vogelstein. 2015. "Variation in Cancer Risk among Tissues Can Be Explained by the Number of Stem Cell Divisions." *Science* 347 (6217): 78–81. https://doi.org/10.1126/science.1260825.

Torii, Manabu, Cecilia N. Arighi, Gang Li, Qinghua Wang, Cathy H. Wu, and K. Vijay-Shanker. 2015. "RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12 (1): 17–29.

https://doi.org/10.1109/TCBB.2014.2372765.

Tourneau, Christophe Le, Jean Pierre Delord, Anthony Gonçalves, Céline Gavoille, Coraline Dubot, Nicolas Isambert, Mario Campone, et al. 2015. “Molecularly Targeted Therapy Based on Tumour Molecular Profiling versus Conventional Therapy for Advanced Cancer (SHIVA): A Multicentre, Open-Label, Proof-of-Concept, Randomised, Controlled Phase 2 Trial.” *The Lancet Oncology* 16 (13): 1324–34. https://doi.org/10.1016/S1470-2045(15)00188-6.

Trabzuni, Daniah, Adaikalavan Ramasamy, Sabaena Imran, Robert Walker, Colin Smith, Michael E. Weale, John Hardy, and Mina Ryten. 2013. “Widespread Sex Differences in Gene Expression and Splicing in the Adult Human Brain.” *Nature Communications* 4. https://doi.org/10.1038/ncomms3771.

Tran, John C., Leonid Zamdborg, Dorothy R. Ahlf, Ji Eun Lee, Adam D. Catherman, Kenneth R. Durbin, Jeremiah D. Tipton, et al. 2011. “Mapping Intact Protein Isoforms in Discovery Mode Using Top-down Proteomics.” *Nature* 480 (7376): 254–58. https://doi.org/10.1038/nature10575.

Tran, Ngoc Hieu, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. 2017. “De Novo Peptide Sequencing by Deep Learning.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (31): 8247–52. https://doi.org/10.1073/pnas.1705691114.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. “Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.” *Nature Protocols* 7 (3): 562–78. https://doi.org/10.1038/nprot.2012.016.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5): 511–15. https://doi.org/10.1038/nbt.1621.

Tsutsui, Shusaku, Reiko Yamamoto, Hiroyasu Iishi, Masaharu Tatsuta, Motomu Tsuji, and Nobuyuki Terada. 1992. “Promoting Effect of Ovariectomy on Hepatocellular Tumorigenesis Induced in Mice by 3′methyl-4-Dimethylaminoazobenzene.” *Virchows Archiv B Cell Pathology Including Molecular Pathology* 62 (1): 371–75. https://doi.org/10.1007/BF02899706.

Tubbs, Anthony, and André Nussenzweig. 2017. “Endogenous DNA Damage as a Source of Genomic Instability in Cancer.” *Cell*. Cell Press. https://doi.org/10.1016/j.cell.2017.01.002.

Tukiainen, Taru, Alexandra Chloé Villani, Angela Yen, Manuel A. Rivas, Jamie L. Marshall,

Rahul Satija, Matt Aguirre, et al. 2017. “Landscape of X Chromosome Inactivation across Human Tissues.” *Nature* 550 (7675): 244–48. https://doi.org/10.1038/nature24265.

Türei, Dénes, Tamás Korcsmáros, and Julio Saez-Rodriguez. 2016. “OmniPath: Guidelines and Gateway for Literature-Curated Signaling Pathway Resources.” *Nature Methods*. Nature Publishing Group. https://doi.org/10.1038/nmeth.4077.

Turro, Ernest, Shu Yi Su, Ângela Gonçalves, Lachlan J.M. Coin, Sylvia Richardson, and Alex Lewin. 2011. “Haplotype and Isoform Specific Expression Estimation Using Multi-Mapping RNA-Seq Reads.” *Genome Biology* 12 (2). https://doi.org/10.1186/gb-2011-12-2-r13.

Tyanova, Stefka, Tikira Temu, and Juergen Cox. 2016. “The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics.” *Nature Protocols* 11 (12): 2301–19. https://doi.org/10.1038/nprot.2016.136.

Ubersax, Jeffrey A., and James E. Ferrell. 2007. “Mechanisms of Specificity in Protein Phosphorylation.” *Nature Reviews Molecular Cell Biology*. Nature Publishing Group. https://doi.org/10.1038/nrm2203.

Udeshi, Namrata D., Tanya Svinkina, Philipp Mertins, Eric Kuhn, D. R. Mani, Jana W. Qiao, and Steven A. Carr. 2013. “Refined Preparation and Use of Anti-Diglycine Remnant (k-ε-Gg) Antibody Enables Routine Quantification of 10,000s of Ubiquitination Sites in Single Proteomics Experiments.” *Molecular and Cellular Proteomics* 12 (3): 825–31. https://doi.org/10.1074/mcp.O112.027094.

Valenzuela-Escárcega, Marco A., Özgün Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T. Morrison. 2018. “Large-Scale Automated Machine Reading Discovers New Cancer-Driving Mechanisms.” *Database* 2018 (2018). https://doi.org/10.1093/database/bay098.

Varga, Tamas, Zsolt Czimmerer, and Laszlo Nagy. 2011. “PPARs Are a Unique Set of Fatty Acid Regulated Transcription Factors Controlling Both Lipid Metabolism and Inflammation.” *Biochimica et Biophysica Acta - Molecular Basis of Disease*. Biochim Biophys Acta. https://doi.org/10.1016/j.bbadis.2011.02.014.

Vasaikar, Suhas, Chen Huang, Xiaojing Wang, Vladislav A. Petyuk, Sara R. Savage, Bo Wen, Yongchao Dou, et al. 2019. “Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities.” *Cell* 177 (4): 1035-1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.

Vavouri, Tanya, Jennifer I. Semple, Rosa Garcia-Verdugo, and Ben Lehner. 2009. “Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity.” *Cell* 138 (1): 198–208. https://doi.org/10.1016/j.cell.2009.04.029.

Vazquez, Miguel, Victor de la Torre, and Alfonso Valencia. 2012. “Chapter 14: Cancer Genome Analysis.” *PLoS Computational Biology* 8 (12): 1–9. https://doi.org/10.1371/journal.pcbi.1002824.

Vigneron, Suzanne, Lena Sundermann, Jean Claude Labbé, Lionel Pintard, Ovidiu Radulescu, Anna Castro, and Thierry Lorca. 2018. “Cyclin A-Cdk1-Dependent Phosphorylation of Bora Is the Triggering Factor Promoting Mitotic Entry.” *Developmental Cell* 45 (5): 637-650.e7. https://doi.org/10.1016/j.devcel.2018.05.005.

Villén, Judit, and Steven P. Gygi. 2008. “The SCX/IMAC Enrichment Approach for Global Phosphorylation Analysis by Mass Spectrometry.” *Nature Protocols* 3 (10): 1638. https://doi.org/10.1038/nprot.2008.150.

Wagner, A. D., S. Oertelt-Prigione, A. Adjei, T. Buclin, V. Cristina, C. Csajka, G. Coukos, et al. 2019. “Gender Medicine and Oncology: Report and Consensus of an ESMO Workshop.” *Annals of Oncology*. Oxford University Press. https://doi.org/10.1093/annonc/mdz414.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. 2012. “Measurement of MRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent among Samples.” *Theory in Biosciences* 131 (4): 281–85. https://doi.org/10.1007/s12064-012-0162-3.

Wagner, Sebastian A., Petra Beli, Brian T. Weinert, Michael L. Nielsen, Jürgen Cox, Matthias Mann, and Chunaram Choudhary. 2011. “A Proteome-Wide, Quantitative Survey of In Vivo Ubiquitylation Sites Reveals Widespread Regulatory Roles.” *Molecular & Cellular Proteomics* 10 (10): M111.013284. https://doi.org/10.1074/mcp.m111.013284.

Wagner, Susan, Anna Herrmannová, Radek Malík, Lucie Peclinovská, and Leoš Shivaya Valášek. 2014. “Functional and Biochemical Characterization of Human Eukaryotic Translation Initiation Factor 3 in Living Cells.” *Molecular and Cellular Biology* 34 (16): 3041–52. https://doi.org/10.1128/mcb.00663-14.

Walport, Louise J., Richard J. Hopkinson, Melanie Vollmar, Sarah K. Madden, Carina Gileadi, Udo Oppermann, Christopher J. Schofield, and Catrine Johansson. 2014. “Human UTY(KDM6C) Is a Male-Specific N∈-Methyl Lysyl Demethylase.” *Journal of Biological Chemistry* 289 (26): 18302–13. https://doi.org/10.1074/jbc.M114.555052.

Walsh, Christopher T., Sylvie Garneau-Tsodikova, and Gregory J. Gatto. 2005. “Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications.” *Angewandte Chemie - International Edition*. Angew Chem Int Ed Engl. https://doi.org/10.1002/anie.200501023.

Wang, Huanan, Guangxu Deng, Meiling Ai, Zhijun Xu, Tingyu Mou, Jiang Yu, Hao Liu, Shuang Wang, and Guoxin Li. 2019. “Hsp90ab1 Stabilizes LRP5 to Promote Epithelial–Mesenchymal Transition via Activating of AKT and Wnt/β-Catenin Signaling

Pathways in Gastric Cancer Progression." *Oncogene* 38 (9): 1489–1507. https://doi.org/10.1038/s41388-018-0532-5.

Wang, Jing, Zihao Ma, Steven A. Carr, Philipp Mertins, Hui Zhang, Zhen Zhang, Daniel W. Chan, et al. 2017. "Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction." *Molecular & Cellular Proteomics* 16 (1): 121–34. https://doi.org/10.1074/mcp.M116.060301.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. "ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data." *Nucleic Acids Research* 38 (16). https://doi.org/10.1093/nar/gkq603.

Wang, Liang Bo, Alla Karpova, Marina A. Gritsenko, Jennifer E. Kyle, Song Cao, Yize Li, Dmitry Rykunov, et al. 2021. "Proteogenomic and Metabolomic Characterization of Human Glioblastoma." *Cancer Cell* 39 (4): 509-528.e20. https://doi.org/10.1016/j.ccell.2021.01.006.

Wang, Man Tzu, Matthew Holderfield, Jacqueline Galeas, Reyno Delrosario, Minh D. To, Allan Balmain, and Frank McCormick. 2015. "K-Ras Promotes Tumorigenicity through Suppression of Non-Canonical Wnt Signaling." *Cell* 163 (5): 1237–51. https://doi.org/10.1016/j.cell.2015.10.041.

Wang, Ouchen, Zhouci Zheng, Qingxuan Wang, Yixiang Jin, Wenxu Jin, Yinghao Wang, Endong Chen, and Xiaohua Zhang. 2017. "ZCCHC12, a Novel Oncogene in Papillary Thyroid Cancer." *Journal of Cancer Research and Clinical Oncology* 143 (9): 1679–86. https://doi.org/10.1007/s00432-017-2414-6.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews. Genetics* 10 (1): 57–63. https://doi.org/10.1038/nrg2484.

Watt, Stephen A., Karin J. Purdie, Nicoline Y. Den Breems, Michelle Dimon, Sarah T. Arron, Angela T. McHugh, Dylan J. Xue, et al. 2015. "Novel CARD11 Mutations in Human Cutaneous Squamous Cell Carcinoma Lead to Aberrant NF-KB Regulation." *American Journal of Pathology* 185 (9): 2354–63. https://doi.org/10.1016/j.ajpath.2015.05.018.

Wei, Wei, Young Shik Shin, Min Xue, Tomoo Matsutani, Kenta Masui, Huijun Yang, Shiro Ikegami, et al. 2016. "Single-Cell Phosphoproteomics Resolves Adaptive Signaling Dynamics and Informs Targeted Combination Therapy in Glioblastoma." *Cancer Cell* 29 (4): 563–73. https://doi.org/10.1016/j.ccell.2016.03.012.

Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R.Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Chris Sander, et al. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature Genetics*. Nature Publishing Group. https://doi.org/10.1038/ng.2764.

Wells, Jonathan N., L. Therese Bergendahl, and Joseph A. Marsh. 2016. "Operon Gene

Order Is Optimized for Ordered Protein Complex Assembly." *Cell Reports* 14 (4): 679–85. https://doi.org/10.1016/j.celrep.2015.12.085.

Wiese, Sebastian, Kai A. Reidegeld, Helmut E. Meyer, and Bettina Warscheid. 2007. "Protein Labeling by ITRAQ: A New Tool for Quantitative Mass Spectrometry in Proteome Research." *Proteomics* 7 (3): 340–50. https://doi.org/10.1002/pmic.200600422.

Witze, Eric S., William M. Old, Katheryn A. Resing, and Natalie G. Ahn. 2007. "Mapping Protein Post-Translational Modifications with Mass Spectrometry." *Nature Methods*. Nat Methods. https://doi.org/10.1038/nmeth1100.

Wollam, Joshua, and Adam Antebi. 2011. "Sterol Regulation of Metabolism, Homeostasis, and Development." *Annual Review of Biochemistry* 80 (July): 885–916. https://doi.org/10.1146/annurev-biochem-081308-165917.

Wu, Linfeng, Sophie I. Candille, Yoonha Choi, Dan Xie, Lihua Jiang, Jennifer Li-Pook-Than, Hua Tang, and Michael Snyder. 2013. "Variation and Genetic Control of Protein Abundance in Humans." *Nature* 499 (7456): 79–82. https://doi.org/10.1038/nature12223.

Wu, Ming Heng, Wen Lung Ma, Cheng Lung Hsu, Yuh Ling Chen, Jing Hsiung James Ou, Charlotte Kathryn Ryan, Yao Ching Hung, Shuyuan Yeh, and Chawnshang Chang. 2010. "Androgen Receptor Promotes Hepatitis B Virus-Induced Hepatocarcinogenesis through Modulation of Hepatitis B Virus RNA Transcription." *Science Translational Medicine* 2 (32). https://doi.org/10.1126/scitranslmed.3001143.

Wu, Ronghu, Noah Dephoure, Wilhelm Haas, Edward L. Huttlin, Bo Zhai, Mathew E. Sowa, and Steven P. Gygi. 2011. "Correct Interpretation of Comprehensive Phosphorylation Dynamics Requires Normalization by Protein Expression Changes." *Molecular and Cellular Proteomics* 10 (8). https://doi.org/10.1074/mcp.M111.009654.

Xie, Shao Hua, and Jesper Lagergren. 2016. "A Global Assessment of the Male Predominance in Esophageal Adenocarcinoma." *Oncotarget* 7 (25): 38876–83. https://doi.org/10.18632/oncotarget.9113.

Xie, Yinlong, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, et al. 2014. "SOAPdenovo-Trans: De Novo Transcriptome Assembly with Short RNA-Seq Reads." *Bioinformatics* 30 (12): 1660–66. https://doi.org/10.1093/bioinformatics/btu077.

Xu, Yue, D. Eric Anderson, and Yihong Ye. 2016. "The HECT Domain Ubiquitin Ligase HUWE1 Targets Unassembled Soluble Proteins for Degradation." *Cell Discovery* 2 (November). https://doi.org/10.1038/celldisc.2016.40.

Yadav, Satya P. 2007. "The Wholeness in Suffix -Omics, -Omes, and the Word Om." *Journal of Biomolecular Techniques*. The Association of Biomolecular Resource

Facilities. /pmc/articles/PMC2392988/.

Yaffe, Michael B. 2019. “Why Geneticists Stole Cancer Research Even Though Cancer Is Primarily a Signaling Disease.” *Science Signaling* 12 (565). https://doi.org/10.1126/scisignal.aaw3483.

Yamamoto, Reiko, Hiroyasu Iishi, Masaharu Tatsuta, Motomu Tsuji, and Nobuyuki Terada. 1991. “Roles of Ovaries and Testes in Hepatocellular Tumorigenesis Induced in Mice by 3′-methyl-4-dimethylaminoazobenzene.” *International Journal of Cancer* 49 (1): 83–88. https://doi.org/10.1002/ijc.2910490116.

Yang, Hui, Dan Ye, Kun Liang Guan, and Yue Xiong. 2012. “IDH1 and IDH2 Mutations in Tumorigenesis: Mechanistic Insights and Clinical Perspectives.” *Clinical Cancer Research*. Clin Cancer Res. https://doi.org/10.1158/1078-0432.CCR-12-1773.

Yang, Jing, Ji Nie, Xuelei Ma, Yuquan Wei, Yong Peng, and Xiawei Wei. 2019. “Targeting PI3K in Cancer: Mechanisms and Advances in Clinical Trials.” *Molecular Cancer*. BioMed Central Ltd. https://doi.org/10.1186/s12943-019-0954-x.

Yang, Xiaoyu, Vijay Dondeti, Rebecca Dezube, Dawn M. Maynard, Lewis Y. Geer, Jonathan Epstein, Xiongfong Chen, Sanford P. Markey, and Jeffrey A. Kowalak. 2004. “DBParser: Web-Based Software for Shotgun Proteomic Data Analyses.” *Journal of Proteome Research* 3 (5): 1002–8. https://doi.org/10.1021/pr049920x.

Yao, Reina, Connie G. Chiu, Scott S. Strugnell, Sabrina Gill, and Sam M. Wiseman. 2011. “Gender Differences in Thyroid Cancer: A Critical Review.” *Expert Review of Endocrinology and Metabolism*. Expert Rev Endocrinol Metab. https://doi.org/10.1586/eem.11.9.

Yoshihara, Kosuke, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, et al. 2013. “Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data.” *Nature Communications* 4. https://doi.org/10.1038/ncomms3612.

Young, L., R. L. Jernigan, and D. G. Covell. 1994. “A Role for Surface Hydrophobicity in Protein-protein Recognition.” *Protein Science* 3 (5): 717–29. https://doi.org/10.1002/pro.5560030501.

Yu, Cheng Ping, Jar Yi Ho, Yi Ting Huang, Tai Lung Cha, Guang Huan Sun, Dah Shyong Yu, Fung Wei Chang, Shu Pin Chen, and Ren Jun Hsu. 2013. “Estrogen Inhibits Renal Cell Carcinoma Cell Progression through Estrogen Receptor-β Activation.” *PLoS ONE* 8 (2). https://doi.org/10.1371/journal.pone.0056667.

Yu, Guangchuang, Li Gen Wang, Yanyan Han, and Qing Yu He. 2012. “ClusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters.” *OMICS A Journal of Integrative Biology* 16 (5): 284–87. https://doi.org/10.1089/omi.2011.0118.

Yu, Guangchuang, Li Gen Wang, Guang Rong Yan, and Qing Yu He. 2015. “DOSE: An

R/Bioconductor Package for Disease Ontology Semantic and Enrichment Analysis." *Bioinformatics* 31 (4): 608–9. https://doi.org/10.1093/bioinformatics/btu684.

Yu, Min, Zheng Tang, Fandi Meng, Minghui Tai, Jingyao Zhang, Ruitao Wang, Chang Liu, and Qifei Wu. 2016. "Elevated Expression of FoxM1 Promotes the Tumor Cell Proliferation in Hepatocellular Carcinoma." *Tumor Biology* 37 (1): 1289–97. https://doi.org/10.1007/s13277-015-3436-9.

Yuan, Yuan, Lingxiang Liu, Hu Chen, Yumeng Wang, Yanxun Xu, Huzhang Mao, Jun Li, et al. 2016. "Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients." *Cancer Cell* 29 (5): 711–22. https://doi.org/10.1016/j.ccell.2016.04.001.

Zahm, Shelia Hoar, and Aaron Blair. 2003. "Occupational Cancer among Women: Where Have We Been and Where Are We Going?" In *American Journal of Industrial Medicine*, 44:565–75. Am J Ind Med. https://doi.org/10.1002/ajim.10270.

Zhang, Bing, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C. Chambers, et al. 2014. "Proteogenomic Characterization of Human Colon and Rectal Cancer." *Nature* 513 (7518): 382–87. https://doi.org/10.1038/nature13438.

Zhang, Bing, Jeffrey R. Whiteaker, Andrew N. Hoofnagle, Geoffrey S. Baird, Karin D. Rodland, and Amanda G. Paulovich. 2019. "Clinical Potential of Mass Spectrometry-Based Proteogenomics." *Nature Reviews Clinical Oncology*. Nature Publishing Group. https://doi.org/10.1038/s41571-018-0135-7.

Zhang, Hui, Tao Liu, Zhen Zhang, Samuel H. Payne, Bai Zhang, Jason E. McDermott, Jian-Ying Zhou, et al. 2016. "Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer." *Cell* 166 (3): 755–65. https://doi.org/10.1016/j.cell.2016.05.069.

Zhang, Ling, Hui Li, Qing hai Ji, Yong xue Zhu, Zhuo ying Wang, Yu Wang, Cai ping Huang, Qiang Shen, Duan shu Li, and Yi Wu. 2012. "The Clinical Features of Papillary Thyroid Cancer in Hashimoto's Thyroiditis Patients from an Area with a High Prevalence of Hashimoto's Disease." *BMC Cancer* 12 (December). https://doi.org/10.1186/1471-2407-12-610.

Zhang, Wei, Ana Bojorquez-Gomez, Daniel Ortiz Velez, Guorong Xu, Kyle S. Sanchez, John Paul Shen, Kevin Chen, et al. 2018. "A Global Transcriptional Network Connecting Noncoding Mutations to Changes in Tumor Gene Expression." *Nature Genetics* 50 (4): 613–20. https://doi.org/10.1038/s41588-018-0091-2.

Zhao, Min, Jingchun Sun, and Zhongming Zhao. 2013. "TSGene: A Web Resource for Tumor Suppressor Genes." *Nucleic Acids Research* 41 (D1). https://doi.org/10.1093/nar/gks937.

Zheng, Jian Feng, Lin Lin Li, Juan Lu, Kun Yan, Wu Hua Guo, and Ji Xiang Zhang. 2015.

"XPD Functions as a Tumor Suppressor and Dysregulates Autophagy in Cultured HepG2 Cells." *Medical Science Monitor* 21 (May): 1562–68. https://doi.org/10.12659/MSM.894303.

Zheng, Kang, Francisco Javier Cubero, and Yulia A. Nevzorova. 2017. "C-MYC-Making Liver Sick: Role of c-MYC in Hepatic Cell Function, Homeostasis and Disease." *Genes* 8 (4): 2–20. https://doi.org/10.3390/genes8040123.