Universiteit Leiden

Master Computer Science

Identifying a Novel Class of Cancer Dependencies:

Paralog Genes Interacting with Common Essentials

Name:	Hermes A. J. Spaink
Student ID:	s1692089
Date:	26/08/2020
Specialisation:	Bioinformatics
1st supervisor:	Katy J. Wolstencroft, PhD
2nd supervisors:	Neekesh V. Dharia, MD, PhD
	Kenneth N. Ross, PhD
	Prof. Kimberly Stegmaier, MD

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Preface

The story behind how this project came about is quite an elaborate one which I will not include here in full. I would, on the other hand, like to thank all the people directly or indirectly involved in making this research and experience possible. Almost one and a half years ago I sent my first email to Neekesh, who I was brought in contact with via Tom Look. After occasionally emailing back and forth for nearly half a year I got to meet him and Ken in person. Already then, about six months before I would eventually move to Boston to start this project, they inspired me with their research.

Although the internship of eight months that I would initially stay for in Boston was cut short to about two and a half months, due to the COVID-19 outbreak, it was still a great experience abroad. I would really like to thank Tom and Shuning for ensuring I had an amazing stay in Boston, along with my housemates and all the great people in the lab of Kim, especially the lab technicians, Amanda, Jerrel, Allie and Caroline, whom I shared an office with.

The research behind this thesis could fortunately go on, though. Neekesh and Ken were great mentors for me during the whole period and I want to thank them for all the time and effort they put into supervising me. Much gratitude also goes to Kim for giving me the opportunity to work and learn in her lab. I also want to thank Clare who provided for great discussions and feedback on the biological side and for using the data on *NXT1* in neuroblastoma which she and collegues generated. And, finally, many thanks go to Katy for her great supervision and being very supportive, yet critical.

Abstract

Cancer remains a leading cause of morbidity and mortality in the world despite advances in understanding its biology and treatment. Therefore, it is essential to continue to investigate this diverse group of diseases and discover new therapeutic targets. Here we present a novel list of potential therapeutic genetic targets in cancer where we focused on genes which are essential in very specific cancer types. Paralog genes are not essential in most cell types due to the buffering factor of their duplicate. We combined the genome-scale cancer dependency map (DepMap), numerous protein-protein interaction databases, several paralog databases and various datasets of gene expression in cancer to propose a list of 429 paralog genes interacting with common essentials as highly selective dependencies. For each of these candidate paralogs, we developed a set of features by which 20 candidate genetic dependencies were selected for further investigation.

One of the higher scoring genes in our selection is the nuclear RNA export factor *NXT1*, which is essential in multiple neuroblastoma cell lines. The behavior of *NXT1* in the *MYCN*-amplified neuroblastoma cell line Kelly is investigated in more detail by performing ChIPseq data analysis. The chromosomal binding locations of NXT1 are investigated to recover the gene types it is most present at. Gene set enrichment analysis showed that NXT1 bound at distal intergenic regions are important for many gene sets related to neuron development. Moreover, the correspondence with the core transcriptional regulatory circuit (CRC), a main driving force behind neuroblastoma survival, is evaluated where we found a large overlap with MYCN binding locations especially at promoter regions. Finally, the effect of NXT1 binding on gene expression is investigated based on RNAseq and protein mass spectrometry. The proteomics data indicated that genes where NXT1 and all the CRC members bind at distal intergenic regions are repressed the most when NXT1 is depleted using a dTAG-13 degradation system. This gene group is also the most enriched for neuroblastoma related gene sets which could explain why *NXT1* plays a crucial role for neuroblastoma. These results shows that our proposed method works well for identification of potential therapeutic targets in cancers.

Contents

Preface i

Abstract iii

1 Introduction 1

1.1 Motivation and research questions . . . 2

1.2 Thesis outline . . . 3

2 Materials & methods 5

2.1 Overview of the data . . . 5

2.2 Experimental setup: novel gene selection . . . 12

2.3 Experimental setup: *NXT1* in neuroblastoma in-depth analysis . . . 19

3 Results: novel gene selection 21

3.1 STRING interaction score . . . 21

3.2 Candidate gene selection . . . 21

3.3 Feature creation . . . 23

3.4 Attribute scoring . . . 27

3.5 Selection of genes of interest . . . 29

4 Results: *NXT1* in Neuroblastoma 31

4.1 *NXT1* peak analysis . . . 31

4.2 Overlap with CRC . . . 32

4.3 Gene expression effect . . . 36

5 Discussion 39

5.1 Paralog genes interacting with common essentials . . . 39

5.2 *NXT1* in neuroblastoma . . . 42

6 Conclusion 43

6.1 Future work . . . 44

Bibliography 45

Appendices 49

A List of cancer types 49

B Treehouse to DepMap disease name mapping 51

C GSEA of enriched vs induced genes 55

Chapter 1

Introduction

Cancer is a group of often life-threatening diseases that severely affect people of any age. Although each cancer type is a unique disease in many forms, all cancers share the property of abnormal cell growth and a disturbed cell cycle caused by a genetic or epigentic changes [1, 2]. Cancerous tumors can go into metastasis and spread throughout the body, which is often fatal. The aim is therefore to apply a treatment before this occurs. On the other hand, in many pediatric cancers this is not the case, instead, here the thought is that this metastatic disease is caused by different biology instead of the fact that it was not caught in time [3].

One possibility for cancer treatment is to therapeutically target specific genes crucial for cancer development, maintenance and growth. To systematically identify genetic vulnerabilities in a certain cancer types, the Broad Institute[1] has been developing a genome-scale CRISPR-Cas9 cancer dependency map, called DepMap [4], over the past few years. This resulted in large datasets of dependency scores for over eighteen thousand genes in hundreds of cancer cell lines. Dependency scores were determined through CRISPR-Cas9 loss-of-function screens and in some datasets through RNA interference (RNAi) screens where the effect of gene knock-out on cell vitality is measured [5, 6].

During a CRISPR-Cas9 loss-of-function screen, a CRISPR-Cas9 genome editing system is coupled with a library of single guide RNAs (sgRNAs) [7, 8]. These sgRNAs direct the CRISPR construct to specific places in the genome to make a double-strand DNA break, which due to the highly error-prone automatic DNA repair mechanism in the cell causes the disruption or knock-out of a gene. These sgRNA libraries are designed carefully to target every gene in the genome at specific loci to result in a gene knock-out. Thus CRISPR-Cas9 loss-of-function screens provide a powerful method for performing genome-scale multiplexed screening through its relative ease and high precision [9].

As there are millions of data points in DepMap, a problem that arises is the identification of potential therapeutic targets. A key property for a drug is to have a therapeutic window, i.e. kill the disease cells whilst sparing normal cells. While this cannot be directly inferred from a genetic screen, genetic targets

[1] `https://broadinstitute.org`

with high selectivity for particular cancer types and not others are often prioritized. In particular, genes which are essential for the proliferation of all cancer cell lines (termed common- or pan-essential genes) are most likely also crucial for other normal cells [10].

A class of genes which may have such unique vulnerabilities are paralog genes. Paralog genes are a particular class of homologous genes with one or more copies in the genome. These copies are the result of gene duplication but can have diverged to slightly different biological functions [11]. Paralog genes play an important role in biology as their similarity can result in different genes with parallel functions, and often, paralog genes can replace each other's function. In the context of cancer biology and genetic dependency this means paralog genes are less likely to be essential than non-paralog genes due to this buffering factor [12].

However, due to differing genetic backgrounds the expression of paralogs is not equal in every cell type which can lead to a significantly different dependency on the other gene [12]. As such, the sensitivity of a cell to the inhibition of a gene (i.e. its essentiality) is strongly linked to the mutation of another gene [13]. This property can cause a unique genetic dependency for certain genes in specific cancer types where their paralog is underexpressed due to transcriptional regulation or a deleterious mutation. Additionally, because of the genetic alterations in cancer the expression or copy number of paralogs or its transcriptional regulators may vary which is something we can leverage [14, 15].

Multiple studies have indicated that genes with a paralog are generally less essential due to the functional compensation effect of the paralog demonstrating the buffering effect of paralogs genes [16–18]. However, in cancer cells, inactivation through deleterious mutations or transcriptional deregulation of the paralog can lead to significant dependency on a gene [12, 14]. In addition, synthetic deactivation of a paralog is shown to have similar effects [19–21]. This suggests that paralog genes could be efficient therapeutic targets in cancers where their paralog is altered.

The nuclear RNA export factor, *NXT1*, is an example of such a paralog gene. It has the gene *NXT2* as a paralog [22] and its primary function is to transport general mRNA from the nucleus to the cytoplasm by forming a heterodimer with the common essential *NXF1* [23–25]. *NXT1* is selectively essential in neuroblastoma [26] and the properties of this gene can be analyzed using various methods. Genomic binding of the *NXT1* product can be investigated using chromatin immuno-precipitation combined with massively parallel DNA sequencing (ChIPseq). Here the regions where NXT1 binds can be identified which could serve as a marker for the genes which rely on NXF1:NXT1 export.

1.1 Motivation and research questions

As mentioned above, it is a challenge to find potential therapeutic genetic targets from genome-scale dependency screens such as DepMap. Because of the reasons stated above, paralog genes are an interesting group of genes to be further investigated and the unique genomic properties in cancer can be leveraged to find very selectively essential paralog genes in cancer.

Pan-cancer essential genes are critical in for the development of most cancers. By identifying molecular mechanisms containing such a common essential *and* a paralog we hypothesized that novel highly selective cancer treatments can be found. This motivation led to the following problem statement (PS) for this thesis:

PS: To what extend can we define a novel class of dependencies for cancer based on selective essential paralog genes interacting with common essentials?

In order to address this PS we want to develop a method for selection of genes based on information from multiple biological databases. In order to develop an optimal selection pipeline the following research questions (RQs) were defined:

RQ1: What is the best way to select for selective essential paralogs interacting with common essentials?

(a) What features can we define to select the most interesting candidates?

(b) How can these features be used for selection of candidates?

RQ2: What are the properties and cellular functions of the genes in this new group of dependencies?

(a) What are the genomic binding regions of these genes and how do they correspond with cancer development?

(b) What is the effect of these interactions on other cellular processes such as gene expression?

1.2 Thesis outline

In chapter 2 we will give a description of the datasets used in our research and list our methods of analysis. The research in this thesis can be split into two sections. First, methods were developed for defining a list of candidate paralog genes interacting with common essentials and we propose a filtering method to select the top candidate genes for further investigation, the results of this are presented in chapter 3. Next, in chapter 4 the results of a more in-depth study on the gene, *NXT1*, are given, where we investigate the functioning of *NXT1* in *MYCN*-amplified neuroblastoma. All our results will be discussed in the subsequent chapter 5. Finally, we will give an overview of our conclusions and hint to future work in chapter 6.

Chapter 2

Materials & methods

In this chapter, the data and methods used will be described. This will cover all the general properties of the data. In the second part of this chapter, all methods will be described. All code used to generate results is available on github here: `https://github.com/hspaink/novel_cancer_dependencies.git`. References are provided to all public data that has been reused and novel data is available upon request.

2.1 Overview of the data

Cancer research is a quickly evolving field of data science as there are rich datasets available, such as genomics, transcriptomics and proteomics, in cell line models. In this project we made extensive use of such available datasets, and by leveraging and combining large data sets, we were able to exploit existing knowledge and contribute new insights.

The first part of this study is based on published datasets including gene dependency data, protein interaction data, and gene expression data in cancer cell lines. For all of these published datasets we will give a brief summary of their contents and construction below.

The second part of this study uses novel analyses and non-published data generated in this study. Here next-generation sequencing (NGS) data on a specific neuroblastoma cancer cell line is used, namely chromatin immunoprecipitation combined with sequencing (ChIPseq), RNA sequencing (RNAseq) and proteomics data. This data was generated by Clare Malone PhD *et al.* in the pediatric oncology laboratory of Kimberly Stegmaier MD at the Dana-Farber Cancer Institute. An overview of these datasets will be given in subsection 2.1.5.

2.1.1 DepMap datasets

The genome-scale cancer dependency map initiative, DepMap, [4] developed at The Broad Institute contains many datasets of different types, such as CRISPR-Cas9 loss-of-function screening [5, 6], gene expression data [27] and drug effect screens [28]. These datasets are updated every three months,

resulting in four releases per year. All the results in this research were produced using the 20Q2 public release [29]. All raw data can be retrieved from the DepMap website[1].

Multiple datasets were used from DepMap. The most fundamental ones being data resulting from the genome-scale CRISPR-Cas9 and RNA interference (RNAi) screens. This includes the *gene effect* and *gene dependency* datasets described in further detail below.

Gene effect and gene dependency

Both the gene effect as the gene dependency data are constituted of a large matrix of thousands of genes against several hundred cell lines. For each gene in the dataset there is a score in the respective cell line. Three large scale screening datasets were used in this research: the DepMap CRISPR screens [5, 6], the DepMap RNAi screens [30], and the Sanger CRISPR screens [5, 31, 32].

In the DepMap CRISPR screens the gene effect and dependency data were defined by the results of genome-scale CRISPR-Cas9 knockout screens of 18,119 genes in 769 cell lines. The scores in the *gene effect* dataset refer to CERES [5] effect scores which resemble the relative effect the knockout of a gene has within in a particular cell line, i.e. how essential the gene is for the cell line. This score is calculated from the sgRNA cell line specific and shared effect and then scaled such that the median for nonessential knockout effect is 0 and the median for an essential knockout effect is -1. CERES reduces false positives and corrects for the copy effect [5]. Because sometimes it is more interesting to know how likely it is that knocking-out a gene has a true effect on the viability of the cell instead of asking how strong the effect is, such probability scores were defined in the *gene dependency* dataset. These scores are derived from the effect score and indicate the probability that the given gene has an actual depletion effect on the cell line. Here a value of > 0.5 indicates that the gene is most likely a dependency in that particular cell line [6].

Before CRISPR screening technology existed, RNAi screens were conducted by the Broad Institute. The current available RNAi dependency data is a collection of DEMETER2 [30] essentiality scores calculated from three large-scale RNAi screening datasets: The Broad Institute Project Achilles [4], Novartis Project DRIVE [33], and the Marcotte *et al.* breast cell line dataset [34]. The final dataset provided a combined score for 16,497 genes across 501 cell lines where the scoring types (gene effect and gene dependency) were similarly defined as in the DepMap CRISPR screens.

The third cancer dependency dataset used in this project came from the Sanger CRISPR screens. This data was processed using the same Achilles pipeline as was used for the DepMap CRISPR data, but using the quality control (QC) of the Sanger project [31]. The Sanger QC differs in the way the single guide RNAs (sgRNAs) were selected. In summary this means read counts from the Sanger's project SCORE were used and the final gene effect scores were calculated using the same CERES pipeline. This dataset contained gene effect and gene dependency scores for 17,799 genes across 318 cell lines.

[1] `https://depmap.org`

Common essentials and selective dependencies

From the DepMap CRISPR screens, a list of common- or pan-essential genes is defined which are a dependency in almost all of the screened cancer cell lines. These common- or pan-essentials constitute a list of 2,123 genes and are selected in a data driven way as follows. First, for a given gene its gene effect score is ranked in each cell line, then the cell lines are arranged in order of increasing gene effect score for that gene. When considering the 90^{th} percentile of least depleted cell lines this shows a bimodal distribution of genes (see Figure 12b in [6]). A threshold for defining the common essentials is defined by taking the point of minimum density in this distribution using a Gaussian smoothing kernel with a width of 0.1.

Although these genes are still essential in the 90^{th} percentile of the screened cancer cell lines this does not mean they are a dependency in *just* these cancer lines. Rather these genes generally have functions which are essential in *any* cell type, also healthy cells. Their range of functions is very diverse, but they are often key in very essential cellular processes. This makes targeting these genes particularly difficult as one would expect on-target effects to other healthy cells in a similar manner as to cancer cells.

Many genes are, however, not pan-essential but rather a dependency in select groups of cell lines due to tumor-specific differences in biology. This gives rise to a group of *selective dependencies*. In order to define this group a normality likelihood ratio test (NormLRT) was performed on the gene knockout CERES results [33]. This score indicates whether a gene has a divergent from normal score profile based on the deviance between the normal distribution and the skewed t-distribution. NormLRT represents the likelihood that a gene's effect scores come from a skewed distribution and by convention a cutoff of 100 is chosen to define a selective dependency.

Omics datasets

Other datasets used from DepMap were the Cancer Cell Line Encyclopedia (CCLE) expression and mutation datasets [27]. The first represents RNAseq gene expression data of 19,144 protein coding genes across 1,372 cancer cell lines. For each gene a score is given indicating the expression in that cell line. This score is the Log_2 transformed RSEM [35] value of RNAseq transcript per million (TPM) gene expression.

The mutation dataset which was used contained mutation data for 19,540 genes in 1,754 cancer cell lines in Mutation Annotation Format (MAF). This format contains various features for each mutation, and for the purposes in this research the data is binarized to being a deleterious mutation or not based on the `Variant_Classification` field. The mutations in this dataset are aggregated from different sources. The primary one being the quarterly updated whole exome sequencing (WES) data generated by the Broad where each mutation call is generated via the CGA WES Characterization Pipeline[2]. These WES-based mutation calls are combined with the existing mutation data from previous releases which includes all mutation calls as described by Ghandi *et al.* in [27].

[2]This pipeline is described here:
`https://docs.google.com/document/d/1VO2kX_fgfUdOx3mBS9NjLUWGZu794WbTepBel3cBg08/edit`. Last accessed: 06-14-2020

Table 2.1: Overview of the PPI databases CORUM [37], STRING [38], SIGNOR [39] and HuRI [40] with their number of only human genes/proteins and the total number of interactions.

Database	Genes	Interactions
CORUM	4,473	4,274
STRING	19,257	11,759,454
SIGNOR	4,294	23,145
HuRI	8,275	52,569

DepMap prediction data

Although not publicly available in the current DepMap releases, an additional internal dataset from DepMap was used. Using several CCLE [27] datasets including gene expression, mutation, methylation, metabolomics and protein level datasets, DepMap conducts a predictive feature search to identify potential biomarkers of dependency. The CCLE data is combined with meta-data such as cell line lineage, histology and disease subtypes, and Cas9 activity, culture type, media conditions and strictly standardized median difference from Project Achilles [5, 6]. This data provides a list of other genes which are associated with each genetic dependency included in DepMap and the Pearson correlation value of the feature association. Dempster *et al.* provide a detailed description on how this model is run [36].

2.1.2 Protein-protein interaction datasets

As one of the main aims in this study was to evaluate whether there exists an interaction between two genes, or rather their protein products, the second major data source consists of all protein-protein interaction (PPI) datasets [37–40]. There are multiple sources of PPI data, all constructed with different levels of evidence, error and coverage. Since no database is ever completely comprehensive and does not contain interactions present in others, as is also shown in Figure 3.3a in the Results, a combination is necessary to achieve an exhaustive search space and minimize false negatives. A caveat with simply combining data from all the datasets is that each one contains false positives with interactions that may not be true. Subsequent filtering is therefore essential in order to get to a biologically representative result. False positive interactions are thus less of an issue in our case, since consecutive added features will reduce such found genes. Therefore, a union of the following filtered PPI databases was taken to grasp the larger picture of all genetic interactions in human cells.

In total four large PPI datasets were included in this study. The number of unique human genes and the interactions in the databases is summarized in Table 2.1.

CORUM

CORUM [37] is in essence not an interaction dataset but rather a manually curated database of mammalian protein complexes. It includes a list of mainly human (67%), mouse (15%) and rat (10%) experimentally verified protein complexes. All information is collected from published individual experiments. In total 4,274 complexes are present in the most recent CORUM 3.0 database. For each

protein complex the individual sub-units are provided in various formats such as UniProt IDs and Entrez IDs.

STRING

The STRING dataset [38] is a large, frequently used database of known and predicted PPIs. Both physical and functional interactions are included and are sourced from experimental data, computational prediction, orthology between organisms, and interactions accumulated from other databases. STRING combines all this information to form a large PPI network where for each connection one or multiple scores between 0 and 1000 are given indicating the confidence of the interaction in the respective category. The various scoring fields are: gene neighborhood, gene fusion, gene cooccurrence, gene coexpression, experiments/biochemistry, annotated pathways, and textmining. For some scoring category also a *transferred* score is given. This score is computed from interactions in a different organism and then transferred via homology or orthology. In total STRING contains over three billion interactions across over five thousand organisms.

SIGNOR

The SIGnaling Network Open Resource, SIGNOR [39], aggregates published human, mouse and rat signaling information as binary causal relationships between biological substances. For each interaction it stores the direction, type and effect of the reaction. In total almost 23,000 manually-annotated causal relationships are stored between mostly proteins but also other biological entities such as chemicals, phenotypes and complexes are enclosed.

HuRI

The HuRI database [40] is a human-only protein interactome map. It contains about 53,000 binary undirected PPIs. These PPIs were accumulated from several yeast two-hybrid (Y2H) screenings of 2,000 by 2,000 genes and combining these with known systematic functional screens. The novelty of the HuRI database causes it to contain many new and previously not documented PPIs as is also visualized in Figure 3.3a.

2.1.3 Paralog genes datasets

The third key criteria is whether a gene has a paralog. The following data type describes paralog genes. For the same reason that not one database is completely comprehensive, two datasets were combined (see also Figure 3.3b in the Results). Again, false positives are not less of an issue since subsequent filtering will emphasize those paralogs with strong buffering effects. Both databases described below contained a list of genes with their paralog copy.

PANTHER

The Protein ANalysis THrough Evolutionary Relationships (PANTHER) [41] database is a system for gene and protein classification. PANTHER contains much more details than paralog and ortholog information alone such as gene families and evolutionary related proteins, molecular functions, the protein function in biological process, and the function within pathways. We use this database to construct a list of near 13,000 paralog genes across several species.

DGD

The Duplicated Genes Database (DGD) [42] is a database for co-located and duplicated genes particularly. It provides paralog information for genes in nine different species. For human specific genes DGD stores about three and a half thousand paralog genes. DGD was constructed through matching of similar genes on a sequence level and enriched with information on gene groups with similar function.

2.1.4 Treehouse expression data

Besides the large expression dataset available from DepMap, the Treehouse primary tumor expression dataset [43] is used in this study. The Treehouse Childhood Cancer Initiative aims to enable the sharing of pediatric cancer genomic data. From many pediatric tumors gene expression data is collected that is combined with publicly available gene expression data including TARGET and TCGA. This results in a cancer-wide (pediatric and adult) gene expression database from over 12,000 samples along with clinical data such as age, gender and disease type. In this study Treehouse version 11 (released April 2020) is used. The TPM Expression dataset contains log2-normalized TPM gene expression values for 18,119 genes in 12,747 tumor samples. The Clinical Data contains metadata for each of these samples including the disease type.

2.1.5 Neuroblastoma data

The effect and function of the paralog gene *NXT1* in neuroblastoma (NB), where it plays an important role, is investigated in further detail using the representative *MYCN*-amplified NB cell line Kelly. Data for this cancer cell line were used to study specific aspects of *NXT1* biology. These datasets are not publicly available and were obtained through internal communication.

To investigate the function of *NXT1* a genetically modified cell line of Kelly was generated in which endogenous *NXT1* is knocked out and exogenous degron-tagged *NXT1* is expressed. This cell line was treated with two different conditions, either with DMSO (which is the vehicle and should not impact NXT1 levels) or dTAG-13 to specifically deplete and remove NXT1 [44]. The former will subsequently be referred to as the DMSO line and the latter as the dTAG or degraded line. Thus, we compared the conditions where NXT1 was present and degraded with ChIPseq of NXT1. Additionally, RNAseq and mass spectrometry (MS) data of proteins was used to investigate the effects on genome-wide gene expression.

ChIPseq data

ChIPseq provides a method for analyzing epigenetic chromatin structure-based DNA interactions [45,46]. Through binding of a specific protein to DNA followed by extraction of these regions and subsequent DNA sequencing (DNAseq), the regions in the genome can be located which interact with the protein of interest. During ChIPseq a protein of interest is bound to DNA, the chromatin is then fragmented and the protein bound regions are extracted and sequenced. The sequenced reads are then aligned to a reference genome to recover the binding locations of the protein which are also referred to as peaks due to the sums of the read coverage corresponding to peaks of alignment quality. All peaks then pass a computational quality control step to reduce noise and false positives.

Here, ChIPseq is used to obtain the genomic regions where the protein NXT1 binds to. ChIPseq data was available for both degraded and DMSO conditions. Browser Extensible Data (BED) files containing the binding locations of NXT1 are available resulting from the peak calling in the ChIPseq data processing. In the data, there are 5,576 peaks reported for the DMSO Kelly line. For the dTAG line 163 peaks are reported. For each peak the genomic range is given, i.e. the chromosome, genomic start position and genomic end position.

Besides for NXT1, ChIPseq data was available for the members of the transcriptional core regulatory circuit (CRC) [47] in Kelly. The CRC is a tissue specific driving factor to establish and remain cell state in several cancers including *MYCN*-amplified NB. CRCs are formed by interconnected positive, feed-forward self-regulating loops of transcription factors which are generally marked by heavy acetylation at histone H3-lysine 27 (H3K27ac) [48–50]. The CRC in Kelly includes the genes *GATA3*, *HAND2*, *ISL1*, *MYCN*, *PHOX2B* and *TBX2*. Besides this, Durbin *et al.* [47] also provided ChIPseq data on H3K27ac in Kelly.

RNAseq data

In order to study the effect of *NXT1* degradation on gene expression, RNAseq was performed. RNAseq and data preprocessing was performed for 150 base pair paired-end reads on the Novoseq 6000 sequencing platform. Quality control (QC) was performed and the reads were mapped to the GRCh37.p13/hg19 human genome. Gene level reads were summarized by counting the reads that overlapped the gencode v19 annotated gene exons. Gene counts were then used to quantify differentially expressed genes between the experimental and control conditions using DESeq2 [51]. Instead of focusing mostly on a qualitative expression change, DEseq2 allows for a good quantitative analysis through its shrinkage estimation for dispersions and fold changes. Hence, improving stability and interpretability of the estimates.

In terms of the lab, for the RNAseq experiment everything was performed in triplicate. Cells were plated and then the following day treated with 500nM dTAG13 or DMSO for 6 hours. They were then detached with trypsin and split into two pellets normalized by cell number. Whole cell RNA was extracted from one using the Qiagen RNeasy kit, and cytoplasmic and nuclear RNA were isolated from the other using

the cytoplasmic fractionation modification for the RNeasy Kit.

RNAseq was performed to measure gene expression in different cellular regions, resulting in data for whole cell, nuclear and cytoplasmic gene expression. Here, for the nuclear and cytoplasmic DEseq data all three cell groups were used and for the whole cell data two cell groups as the third was discarded during QC. These datasets were filtered to only contain genes with at least 5 reads in all samples. For the whole cell, nuclear and cytoplasmic measurements DEseq values for respectively 14,976, 15,956 and 14,898 genes were available after filtering. A DEseq value here refers to an adjusted p-value and log_2 shrunken fold change for a gene which were computed using DEseq2 [51].

Proteomics data

The proteomics experiment was similar except here four replicates for each condition were used and cells were treated for 2, 6, or 24 hours. Cells were collected by scraping and then pelleted into low-protein adhering tubes. At the three different time points after degradation of *NXT1*, cells were lysed and total protein was determined using a bicinchoninic acid (BCA) assay. Proteomics data was measured using MS of TMT labeled protein. The acquired data was processed to a have a FDR of 1% for all proteins. Next, log_2 shrunken fold change along with a p-value was determined for each protein in the data. For the time points of 2, 6 and 24 hours of degradation the expression of respectively 4,260, 4,225 and 4,487 gene products were measured.

2.2 Experimental setup: novel gene selection

The first part of this study focused on the identification of paralog genes interacting with common essentials. In order to come to a meaningful group of genes, several selection and evaluation steps were performed. In the following our methods and the implementations of these selections will be described. A diagram of the analysis and filtering pipeline showing where in the selection each database is used is given in Figure 3.2 in the Results, as described in subsection 2.2.2. From the list of *Essential candidates* in Figure 3.2 a further selection is made by filtering on additional features which are described in subsection 2.2.3.

The implementations of the bioinformatics analyses pipelines were done using Python 3 [52] in Jupyter Notebook [53]. The Python library pandas [54, 55] was used for storing and manipulating all datasets.

2.2.1 Data preprocessing

Gene IDs

Although genes are usually referred to in literature using their gene symbol or name (e.g. *NXT1*), for purposes of reproducible analyses a more stable and reproducible method of gene identification is preferable. There exist multiple gene identifier (ID) formats, the most common ones include Ensembl IDs [56], UniProt IDs [57], PubMed Accessions [58] and NCBI Entrez IDs [59]. All these methods

originated from their equally named gene databases. Although HUGO symbols are the globally preferred identification method for genes, each dataset of biological data may use a different version of the symbol. A major problem is that there is not a complete one-to-one mapping between the different IDs, as multiple Ensembl IDs can map to a single Entrez ID for instance [60]. Since this is a common occurring problem in bioinformatics, various tools exist to convert between formats [61–65]. These are mainly web-based and rely on large databases of linked up to date IDs.

Data preprocessing is required on several of the PPI and paralog datasets to convert their gene identifiers to equivalent formats. All DepMap datasets make use of a combination of HUGO gene symbols and Entrez IDs. Because Entrez IDs are consistent and commonly used in many other databases and analysis tools this gene identification system was chosen as the leading type in this research.

The databases STRING, SIGNOR, HuRI and PANTHER make use of either Ensembl protein-, Ensembl gene- or UniProt IDs, however. A mapping for these IDs to Entrez IDs is therefore made. In order to overcome the facts that common conversion tools are not always implementable in a custom script and that their back-end databases might not cover all required genes, custom mappings were constructed which are available in the supplementary data on github.

For Ensembl protein- and gene IDs a frequently updated mapping is available from NCBI. This dataset covers over 2 million unique genes and for each entry stores the Entrez ID, the gene symbol, the Ensembl gene ID and the possible Ensembl protein ID. Because, due to alternative splicing, multiple transcripts are possible from a single gene, one gene may have multiple entries with different Ensemble protein IDs. If for some reason an Ensembl ID cannot be found in this list, though, the Ensembl ID passed through a web query to the NCBI website[3] and the Entrez ID is retrieved from the web page. This method ensures the most up to date ID mappings were used.

Although somewhat tedious, a similar approach was taken for the UniProt to Entrez conversion. However, in this case there was no large data mapping available from NCBI. Therefore, all UniProt IDs were queried to the UniProt website from which the Entrez ID was retrieved.

STRING selection

From the CORUM, SIGNOR and HuRI datasets it was straightforward to select relevant PPIs. From these only human genes were selected for the subsequent analyses. However, the STRING database required a somewhat different approach.

As mentioned in section 2.1.2, the STRING database attaches a series of scores to each of its PPIs. To find a good score cutoff several analyses were performed. For all the PPIs in CORUM the STRING interaction score was retrieved, if contained in STRING. Since all the interactions in CORUM are manually curated (i.e. sourced from scientific publications) this allowed for a good benchmark to investigate the optimal STRING score cutoff.

[3] https://www.ncbi.nlm.nih.gov/

As each scoring metric in STRING refers to a probability of the interaction being true based on the score specific evidence, the scores can be combined. Using a single score allows for easier PPI selection from the dataset. From the seven scores present in STRING, the top three highest scoring metrics s_i for the CORUM PPIs were combined to a single score as follows

$$S_{STRING} = 1 - \prod_{i=1}^{3}(1 - s_i)(1 - s_i^{trans}) \tag{2.1}$$

where s_i^{trans} refers to the *transferred* scoring field of a score in the top 3 scores in STRING. Note, each score in STRING has a value between 0 and 1000 and was first normalized to lie between 0 and 1 through division by 1000. The top scoring metrics where selected by manual inspection of their scores for the CORUM PPI in a distribution plot. As presented in section 3.1 in the Results, the top three scores were the *experiments score*, the *databases score* and the *coexpression score*. These scores, along with their transferred scores were used to constitute the combined score.

A cutoff value of 0.6 was taken to select PPIs from STRING. This somewhat low cutoff allows for a wider search space of PPIs and to, hence, avoid missing possible interesting interactions in the final results. It does incur, however, that stringent subsequent selection is essential to build a final list of candidates.

2.2.2 Candidate gene selection

The first step of gene selection focused on the two main criteria: paralog genes interacting with common essentials. Where the common essentials were used as defined by DepMap. In order to do this all gene-gene interaction pairs from the processed PPI databases were selected that contained a common essential. In some cases a common essential interacts with another common essential. These interactions were removed from the candidate PPI pairs.

Subsequently, only the paralog genes were selected by screening using the paralog databases. This yields the first group of candidate genes.

Selectivity filtering

Secondly, this list of genes was filtered based on selective dependencies. For all candidate genes the left-skewness was determined based on a NormLRT value ≥ 100 and the mean of the effect values being smaller than the median. Here the effect values refer to the gene effect data from either the Achilles, DEMETER or Sanger screens.

Next, the cell lines for which each gene is a dependency were retrieved based on the gene dependency data from either the Achilles, DEMETER or Sanger screens. Here a gene is determined as essential for a cell line if the dependency value is ≥ 0.5.

From this filtering, the baseline group of essential candidates was formed where the base requirements of our problem statement were addressed. Further filtering as described in the next section will consolidate

this group to a list of interesting candidate genes.

2.2.3 Feature annotation

Additional features were created with which we annotate the candidate genes. On these features we filter in order to obtain a biologically sound list of selective potential therapeutic target genes. The following features were introduced:

1. paralog expression correlation,
2. relative dependency,
3. gene dependency enrichment,
4. paralog primary tumor enrichment,
5. interaction between paralogs and common essentials,
6. paralog mutation correlation,
7. DepMap prediction values.

In the subsequent sections we will describe how these were constructed and used to create certain filters.

Paralog expression correlation

The first feature we introduce to each candidate is correlation between the gene's dependency and the expression of its paralogs. Pearson R correlation values were computed using SciPy [66] between the Achilles gene effect data [5] representing the dependency of a cell line on the candidate gene, and the DepMap CCLE gene expression data [27] for each paralog. This results in an *expression correlation* value for each paralog per candidate gene.

Additionally, specific correlation values were determined for specific cancer types. For all the cell lines in the DepMap screens, meta-data is available including lineage sub-type and disease. From this, 43 disease subgroups were specified representing a specific cancer type, a list of which is given in appendix Table A.1. Examples of these categories are *neuroblastoma*, *lung cancer* or *AML*. Subsequently, for every disease in specific the expression correlation is calculated. The rationale behind this is that within a specific cancer type the candidate gene may be following a pattern of low paralog expression – high gene dependency with much less noise.

Relative dependency

The relative dependency is a measure of in how many cell lines of a cancer type the candidate gene is essential. It is calculated as the number of cell lines for a disease in which the gene is a dependency, divided by the total number of cell lines available for that disease.

Gene dependency enrichment

A third feature which is incorporated is whether the dependency is enriched in any specific cancer type. To investigate this, the distribution of Achilles gene effect scores [5] for the cell lines within a cancer group were compared to the distribution of scores for all remaining cell lines. This was tested within only the group of cell lines in which the candidate gene was a dependency. An independent two-sided T-test implemented using SciPy [66], is performed on the two groups to test the null hypothesis that the two samples have values from the same global distribution. When all p-values were calculated, multi-hypothesis testing p-value adjustment was performed using the Python library statsmodels [67]. Benjamini-Hochberg [68] false discovery rate reduction was applied with an α of 0.05. Next, an adjusted p-value cutoff of 0.1 is used to accept or reject the null hypothesis. Although, higher than a customary p-value cutoff of for instance 0.05, this value was chosen to allow more genes to have an enriched cancer type for indication of relevance. As can be seen from our results in Figure 3.9, several enriched diseases fall just over the 0.05 p-value border but none the less provide a suggestion of differentiated *gene effect* scores. Thus, we assumed with 90% confidence that the dependency values for the cell lines for the disease of interest were *enriched* if $p_{adj} < 0.1$ and the median of this group is smaller than the median of all others (since a lower gene effect score indicates a higher dependency on the gene).

Paralog primary tumor expression enrichment

Besides incorporating the DepMap expression data which is based on cell line models, a second cancer gene expression dataset was included that includes primary tumors. The Treehouse dataset [43] compiles primary tumor gene expression data from multiple sources providing a more comprehensive cancer data source. A complication is, however, that it uses a different method of categorizing diseases. Additionally, genetic dependency data is not available for these primary tumor samples. It is therefore not possible to compute correlations in a straight-forward manner as above. For that reason, enrichments were calculated for each paralog of every candidate gene per Treehouse disease type. It is investigated whether the expression of the paralog is lower in the specific cancer types as defined by Treehouse using an independent two-sided T-test, multiple hypothesis p-value adjustment over all p-values like above, and H_0 rejection with 95% confidence (i.e. if $p_{adj} < 0.05$), similar to the method as described above.

In order to link the Treehouse cancer types back to our previously defined, DepMap based disease types a mapping is made of Treehouse diseases to DepMap diseases. This mapping can be found in the Appendix Table B.1. For computing the disease score, a paralog was considered enriched for a DepMap disease if at least one of the Treehouse diseases mapping to that DepMap disease was enriched.

Interaction between paralogs and common essentials

A key biological factor in this study is whether the paralog of a gene can replace its function. This is only possible if the gene product of the paralog can interact with the same common essential as the gene of interest. A key investigation is therefore to check whether an interaction is documented in the PPI databases.

During the candidate gene selection, as described in subsection 2.2.2, a list of common essentials each candidate gene interacts with is kept. Here, for each paralog of a candidate gene, the interaction between the common essentials is analyzed by screening of the PPI databases. This results in two sub-lists of the paralogs for each candidate gene which interact with either *all* or *any* of the common essentials which form a PPI with the gene of interest. The paralogs forming a PPI with at least one common essential were considered the *interacting paralogs*.

Paralog mutation correlation

Another feature is the correlation between paralog mutation and gene dependency. A deleterious mutation in the paralog gene may result in a loss of function and hence a higher essentially of the candidate gene. Pearson R correlations were therefore computed between the Achilles gene effect data [5] for every candidate gene and CCLE mutation data [27] for each of their paralogs. Here a value of 1 was assigned to mutations annotated as *damaging* or *other non-conserving* in the `Variant_annotation` field. A negative correlation value here indicates that the paralog was generally mutated when the candidate gene was more essential.

It was found that the mutation correlation values were generally quite low. To represent the resulting values more significantly, Z-scores (or standard scores) of the correlations were determined. These were computed by subtracting the mean of all correlations from the mutation correlation and dividing by the standard deviation of all correlations. Means and standard deviations were computed using the Python library NumPy [69].

DepMap prediction scores

Finally, for each candidate gene the DepMap genes marked as top features in the best predictive model for dependency were included. This was used to confirm if a paralog gene's data (RNAseq, mutation, etc.) was a top predictive feature for the candidate gene dependency. If a paralog of a candidate gene was present in the predicted highly associated genes, this score was stored as an additional feature.

2.2.4 Significant diseases

The paralog expression correlation values provide accurate information for the essentiality of the candidate gene cancer wide. In many cases, however, a certain gene plays an important role within a specific cancer only. In order to investigate the diseases which were relevant or significant to a candidate gene score representing the effect of the gene in question on a disease was constructed. Relative dependency for disease, gene dependency enrichment, disease specific paralog expression correlation, and primary tumor paralog expression enrichment were combined into a single disease score S for disease d by computing

$$S_d = \frac{l_d}{L_d} + (E_d^{eff} \rightarrow c_e) + r_{max} + (E_d^{TH} \rightarrow c_e) \tag{2.2}$$

where l_d is the number of lines where the candidate gene was essential and L_d the total number of lines for disease d, E_d^{eff} and E_d^{TH} Boolean values for the enrichment in respectively the DepMap gene effect data and Treehouse paralog expression data with the constant $c_e = 0.6$ being added if the enrichment was true, and r_{max} being the maximal expression correlation ($r_{exp,d}$) for all paralogs computed as

$$r_{max} = max(\{r_{exp,d}(p_1), ..., r_{exp,d}(p_n)\}) \tag{2.3}$$

for all the n paralogs p. In order for the relative dependency to not have a too high impact in the total disease score, it was only taken into account when there were at least two cell lines available for the cancer type in total. This is the case for 41 out of the 43 disease categories. The two cancer types with two or less cell lines available were *Adrenal Cancer* and *Teratoma*.

2.2.5 Paralog scoring

Similarly to how a ranking is defined for each disease, the importance of a paralog is given for every candidate gene. In several cases there were multiple paralogs defined for a single candidate gene. In order to provide some data based information on which of these paralogs is the most relevant in the cancer system, a score is introduced for each paralog. This score is defined as the sum of the (possible) DepMap prediction score, the negative Z-score of the mutation correlation (since a mutation correlation smaller than 0 indicates an interesting correlation), the paralog expression correlation, the maximal disease specific paralog expression correlation and a constant value of 1 if the paralog interacts with *any* of the common essentials interacting with the candidate gene.

2.2.6 Identification of genes of interest

From the total list of candidate genes a subset of *genes of interest* was defined. Based on the various biologically relevant features which were defined as described above a selection was made. In Figure 3.13 in the Results a diagram of the genes of interest selection is given. A differentiation was made between two key aspects: selectivity across all cancers, and selectivity for a specific disease type. To address the first group of genes we define one criterion for cancer-wide paralog expression correlation to be greater than the 95^{th} percentile.

To select for cancer specific interesting genes a filtering based on the more intricate disease score was made. From the distribution of maximal disease scores across all candidates, similar to above, the 95^{th} percentile was taken. Genes were considered top candidates if their maximal disease score was in this quantile. A second filter was added to remove very low paralog expression correlations. The lowest 5^{th} percentile of the maximal disease specific paralog expression correlation for the candidates were discarded.

Finally, genes with no interacting paralogs or lying within the 85^{th} percentile of the number of interacting paralogs were disregarded in the list of genes. This, because the aim of the study was to find simple PPI sub-systems and large interaction networks, i.e. a gene with many paralogs interacting with many

common essentials, would be difficult to potentially target therapeutically.

2.3 Experimental setup: *NXT1* in neuroblastoma in-depth analysis

The second part of this study provided a comprehensive analysis of one of the high ranking candidate genes, *NTX1*, in neuroblastoma-specific cell lines. In order to address RQ2 several analyses were performed as described below. The implementations of these analysis were made in R [70] with the tidyverse [71] package *dplyr* used for many data table manipulations.

2.3.1 NXT1 ChIPseq analyses

Peak overlaps

To evaluate the binding locations of NXT1, the ChIPseq data of the DMSO treated Kelly line was compared to the dTAG line. ChIPseq peaks were acquired in the form of BED files, which were loaded into R as `GRanges` objects. Using the R package ChIPpeakAnno [72] ChIPseq peak files can be compared. This was used to analyze peak overlaps, i.e. investigate correspondence between genomic binding locations. These overlaps were then visualized as Venn diagrams.

From the overlaps individual peak sets could be extracted. This allowed for selection of certain peak set intersections. Such as the intersection between NXT1 peaks in the DMSO and degraded samples.

The `findOverlapsOfPeaks` method of ChIPpeakAnno was limited to a maximum of four peaks ranges. In order to analyze overlaps between more than four sets, such as the comparison between NXT1 peaks and CRC binding locations, a different method was used. From the R package ComplexHeatmap [73] the method `make_comb_mat` allowed for finding peak overlaps and intersections between more peak ranges. Here the value function parameter (`value_fun`) was set to use the length of the intersection (i.e. the number of peaks), instead of the default total genetic length. The same package was also used to generate UpSet plots of the intersections and extract specific overlaps such as the intersection between all CRC members and the NXT1 non-promoter peaks.

Peak annotation

The genomic regions corresponding to binding locations of NXT1 were retrieved through annotation of the peak ranges. Annotations were obtained using the R package ChIPseeker [74]. Given a peak range, the method `annotatePeak` provides the corresponding genetic region (e.g. promoter, distal intergenic, etc.) and gene it has effect on. For all analysis a transcriptional start site (TSS) of three Kb in either direction was used. Furthermore, a TxDb object from the R package `TxDb.Hsapiens.UCSC.hg19.knownGene` was provided. This is in principle an interface to a genomic database, in our case version 19 of the human genome. Finally, the parameter `annoDb` was set to `"org.Hs.eg.db"`, which adds gene information including gene symbol, gene name, and Ensembl and Entrez ID.

From the annotated peak ranges specific peak regions can be selected. Filtering on rows containing

"promoter" in the annotation column allowed to separate the NXT1 promoter peaks from the other NXT1 binding locations.

2.3.2 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) is a computational method to label sets of genes with statistical significant associations to pathways and phenotypes, i.e. find the appropriate gene sets in which the given list of genes is statistically over-represented using a Kolmogorov-Smirnov [75]. The Molecular Signature Database (MSigDb) [75,76] provides a collection of annotated gene sets, commonly used in GSEA. Using the R package clusterProfiler [77] version 3.16.0 and its method `enricher`, enrichments using a hypergeometric test in the MSigDb Hallmark [78], MSigDb Curated, and Gene Onthology (GO) [79,80] gene sets were detected. For these gene sets, all subcategories were included in the analysis. To determine enriched gene sets for a list of genes the `enricher` method was called with the `TERM2GENE` parameter set to a data frame containing a mapping of genes to gene set retrieved with `msigdbr` from the R package msigdbr version 7.1.1. All other parameters were kept at default. Thus, this method determines statistically significant categories and corrects this using a Benjamini-Hochberg [68] p-value adjustment.

2.3.3 Gene expression analysis

The relation of NXT1 binding to gene expression was visually investigated using volcano plots to visually provide an overview of the data. Volcano plots were made using the R package EnhancedVolcano [81]. Here the volcano plots used the adjusted p-values from the DEseq RNAseq data and "standard" p-values from the proteomics data, along with log_2 shrunken fold change values. Cutoff values for the p-value and fold change were set based on visual inspection of the data distributions. The genes with significant p-value and fold change (i.e. past the cutoff) were used to determine the ratio of number of induced to number of repressed genes.

Groups were marked within these volcano plots based on for instance genes belonging to different binding locations of NXT1. For each group individually the ratio between number of induced and number of repressed genes was calculated by taking the counts of genes past the p-value and log_2 adjusted fold change cutoff.

Chapter 3

Results: novel gene selection

In this chapter the results from the novel gene selection are presented. First, the results from the candidate gene selection are given. Afterwards, an overview of the added features are shown across the list of genes. The online supplementary results are accessible via the Github repository.

3.1 STRING interaction score

In order to determine the top scoring metrics from the STRING PPI database [38], the STRING scores of the interactions stored in CORUM [37] were analyzed. Only 77% of the interactions in CORUM were present in STRING. For these interactions their respective STRING scores are shown in Figure 3.1. To establish a single biologically representative score, first the lowest scoring metrics were disregarded. This were the `fusion`, `cooccurence`, `homology` and `neighborhood_score` metrics. Second, the `textmining_score` in the literature indicates interaction, but with varying degrees of empirical evidence in different publications. The score may therefore introduce a larger number of false positives and was thus omitted.

This left the `database_score`, `experiments_score` and `coexpression_score`. Their individual coverage was compared with combinations of the metrics. As shown in Figure 3.1, a linear combination score of these metrics (i.e. `database_experiments_coexpression`) resulted in a higher score for the CORUM PPIs. A straight-forward linear combination was deemed to be sufficient as the aim of this score was to provide a reasonable identification of interactions. No other combinatorial methods were applied, therefore.

3.2 Candidate gene selection

From the PPI databases, the paralog databases and common essentials 4,452 genes were selected as candidates in total following the selection diagram in Figure 3.2. The origins of the candidates are shown in Table 3.1 and Figure 3.3. As can be seen in Table 3.1 and Figure 3.3a most of the candidates

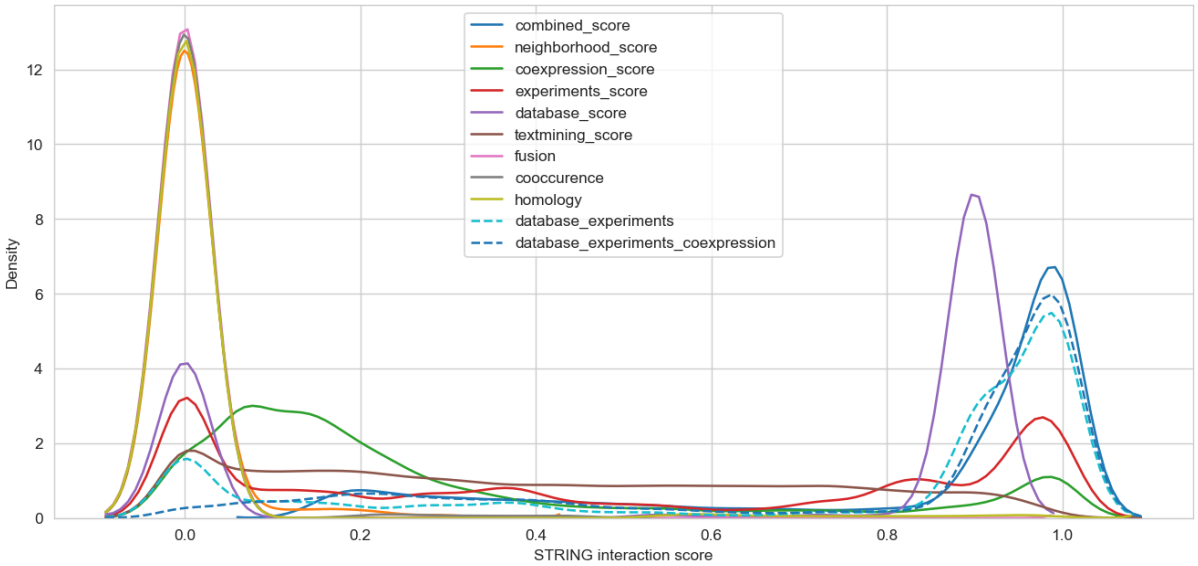


Figure 3.1: Density plot showing distribution of each of the STRING [38] scoring metrics for PPIs in CORUM [37]. The `combined_score` metric is the STRING metric where it combines *all* the scores. The scores with dashed lines (`database_experiments` and `database_experiments_coexpression`) are custom combined scores derived from the respective STRING metrics.

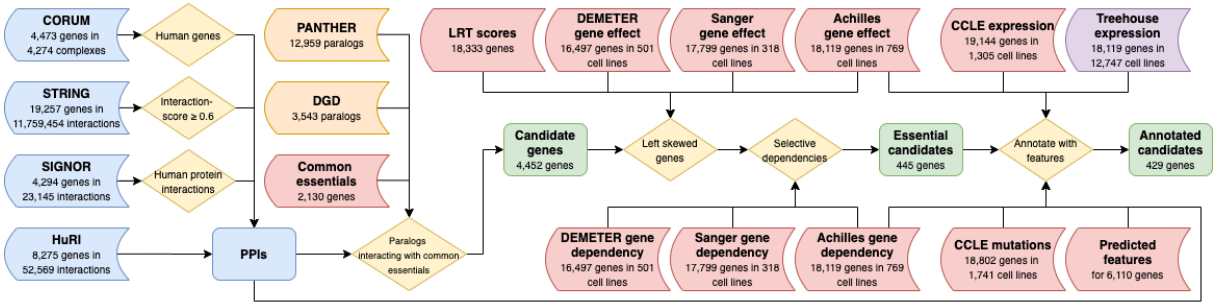


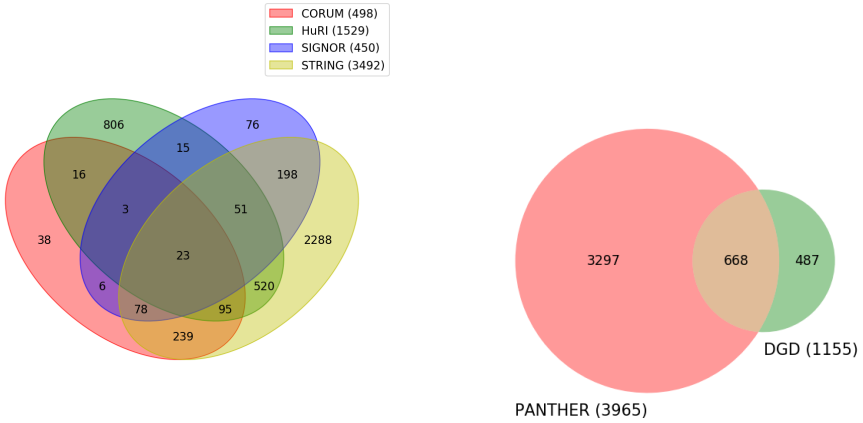
Figure 3.2: Pipeline of candidate gene selection based on filtering from different datasets. In blue are all PPI databases, in orange the paralog databases, in red all DepMap databases and in purple the Treehouse dataset. The yellow diamonds refer to selection or filtering operations. In green intermediary results are shown.

originated from STRING. Additionally, the HuRI [40] dataset also introduces a significant portion of 806 unique candidate genes, which corresponds to 52.7% of all its introduced genes being unique.

Table 3.1: Number of candidate genes as retrieved from the PPI and paralog databases.

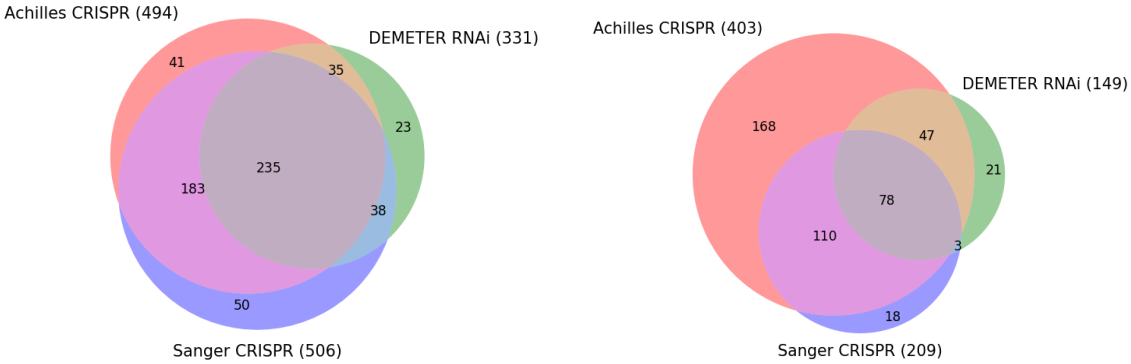
	PANTHER	DGD
CORUM	469	65
STRING	3,126	833
SIGNOR	425	62
HuRI	1,349	451

Filtering for only selective dependencies reduces the number of initially found candidate genes substantially. Selecting based on left skewedness results in 605 candidate genes (see Figure 3.4a). When this list is filtered on being an essential gene in at least one cell line, 445 candidate genes remain (see Figure 3.4b). The genes originate from three dependency screens giving evidence for the essentiality.



(a) PPI databases (b) Paralog databases

Figure 3.3: Venn diagram comparison of origins of the candidate genes as retrieved from the PPI (a) and paralog (b) databases. For each group the number in parentheses shows the total number of genes found from that database.



(a) (b)

Figure 3.4: Comparison of candidate genes filtered on essentiality using three gene dependency screens. In (a) all left skewed genes are shown whereas in (b) only those are shown which are selected on being a dependency in at least one line.

As can be seen in Figure 3.4 the largest part of selective dependencies come from the Achilles CRISPR screen.

3.3 Feature creation

The novel features were introduced and computed for all candidate genes. Due to some features not being able to be available for all candidates (e.g. having too few dependent cell lines for correlations to be calculated), the 445 genes marked as candidates shrunk to 429 candidate genes for the final list. This list of candidate genes is available as a `csv` file in the supplementary materials on GitHub. The results of the custom created features are described below. To illustrate these, in the subsequent examples are shown for the final selection of *most interesting candidates* or the gene *NXT1* in specific.

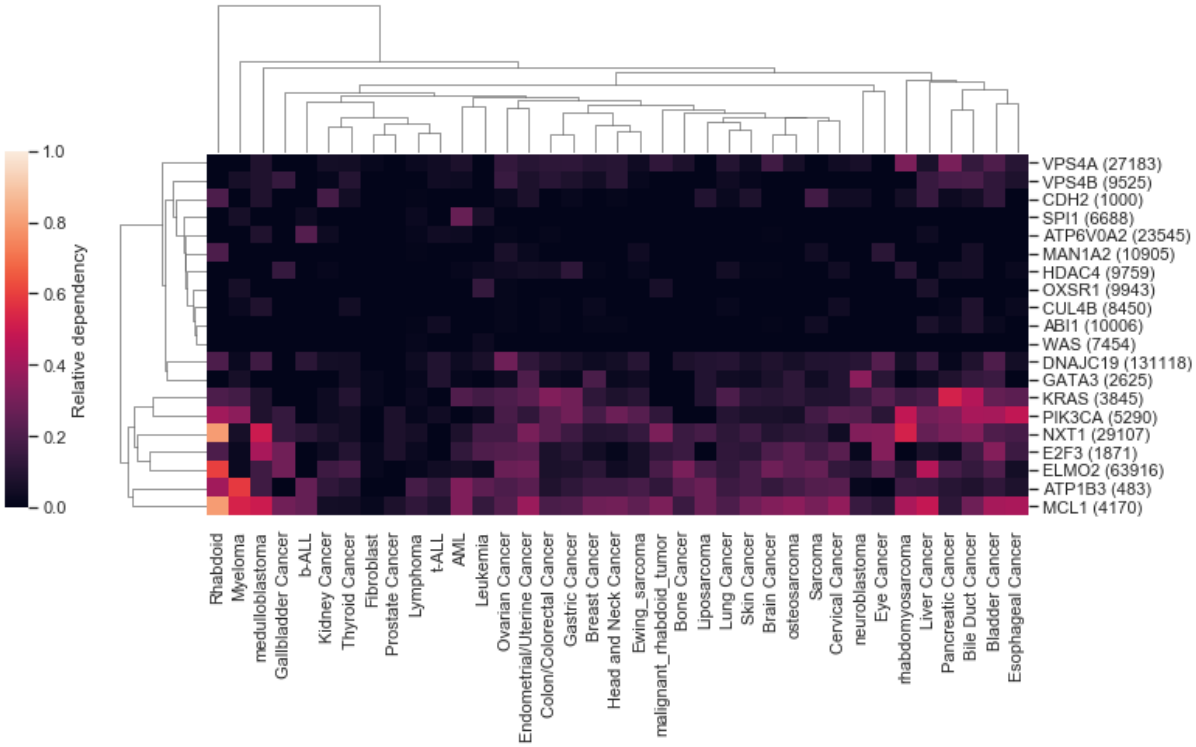
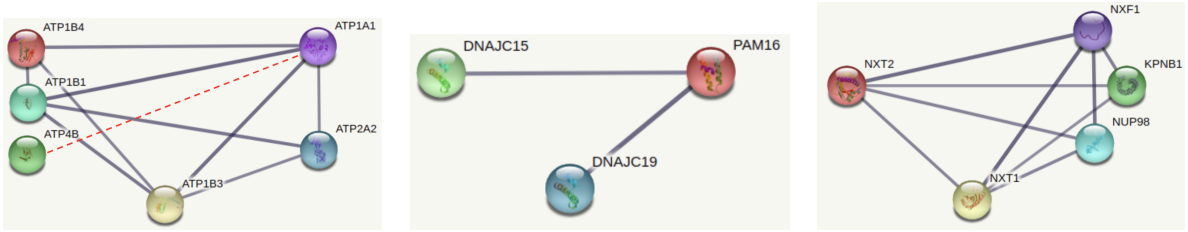


Figure 3.5: Clustering of relative dependency values for the final selection of candidate genes in the different disease groups. Only cancer types where at least one gene had one dependent cell line were included in the plot.



(a) (b) (c)

Figure 3.6: Interaction graphs for several genes in the final selection of candidates. For each graph on the left the interacting paralogs are shown, in the center the candidate gene, on the right the common essentials the candidate and its paralogs interact with. Graphs are retrieved from `https://STRING-db.org` with confidence of experiments, databases, coexpression > 0.600. The red dotted line indicates an interaction found using a different PPI database than STRING.

3.3.1 Relative dependency

From the candidate gene selection, the cell lines in which each candidate was essential were reported. Each cell line was mapped to a discreet cancer type, hence a relative dependency value for each disease could be determined. The relative dependency value was used in computing the *disease score* each candidate gene had for a certain cancer type. The relative dependency of the final selection of most interesting candidates is shown in Figure 3.5, with a maximal value of 1 meaning dependency in all cell lines of the disease. Based on these values, the genes were clustered in hierarchically-clustered heatmap.

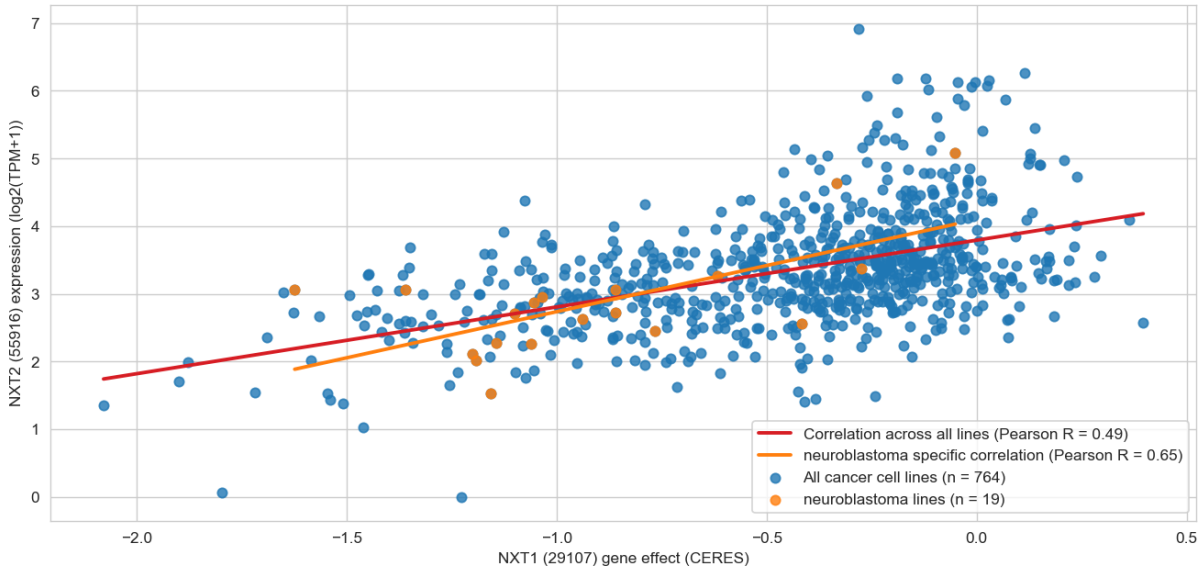


Figure 3.7: *NXT1* dependency against its paralog (*NXT2*) expression scores in 764 cancer cell lines. In red the correlation trend-line is shown. An expression correlation of 0.49 is found across all cell lines. In orange neuroblastoma cell lines are highlighted and the disease specific correlation of 0.65 is drawn as a linear trend-line.

3.3.2 Interaction between paralogs and common essentials

For each candidate gene the interactions between its paralogs and its common essentials were analyzed. Paralogs interacting with at least one of the common essentials were categorized as *interacting paralogs*. These interactions are visualized for three genes in Figure 3.6. Although the candidate genes can have many paralogs and many common essentials, filtering based on internal interactions reduces the pool of relevant ones significantly and hence reveals the biologically interesting genes. The candidate gene *BCL6* for example, interacts with 7 common essentials and has 15 paralogs. Only one of these paralogs (*PATZ1*) interacts with only one of the common essentials (*PFDN5*), however.

For the final selection of candidates, the genes with many interacting paralogs were discarded. This, due to the motivation that the aim of this study is to search for simple PPI subsystems. Hence, candidates falling in the last 15^{th} percentile were disregarded. The 15^{th} percentile in the distribution of number of interacting paralogs for all candidate genes lies at > 5 interacting candidates.

3.3.3 Paralog expression correlation

Correlations between gene effect and paralog expression were computed for all candidate genes and their paralogs. Moreover, correlations within a disease group were defined. An illustration for the gene *NXT1* and its paralog *NXT2* is given in Figure 3.7. Maximums and averages were defined for all the paralogs of a candidate and a differentiation was made for the interacting paralogs as described in subsection 3.3.2.

As shown in Figure 3.8, the gross part of all correlations lies around zero, indicating no clear correlation is found. And the positive expression correlations were quite weak in general due to the large scatter of

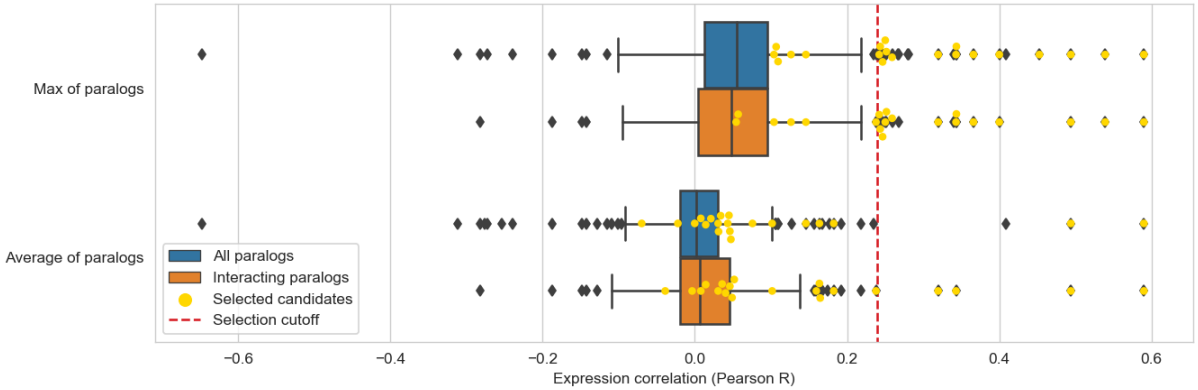


Figure 3.8: Distribution of the found expression correlation values for all candidate genes. In orange only the paralogs interacting with at least one common essential the candidate gene interacts with are included. The *max correlation* refers to the maximal expression correlation values for all or only the interacting paralogs. The *average correlation* refers to the mean of the correlations. The red line indicates the cutoff of the max interacting paralog expression correlation greater than 0.24 for the final selection of interesting candidates. In yellow the finally selected genes of interest are marked.

cell lines. The *max correlation* field shows that a correlation above about 0.2 is the case only for "outlier" candidates. For the final selection a cutoff at the 95^{th} percentile of the max correlation for interacting paralogs was chosen which lies at a correlation value of 0.24 (marked as the red line in Figure 3.8). Hence all candidate genes were selected with higher correlation values.

3.3.4 Gene dependency enrichment

For each candidate gene the disease groups with enriched dependency values were determined. The stringency of enrichments was primarily controlled by the p-value cutoff, which was set to 0.1. This caused 37 out of 429 candidate genes to be enriched for dependency in one or more cancer types. The enriched diseases for the final selection of candidates is shown in Figure 3.9.

3.3.5 Paralog primary tumor enrichment

Because of the non-triviality of calculating gene effect to paralog expression correlations for the Treehouse [43] data, for all the paralogs of the candidate genes enrichments in Treehouse expression were determined. Compared to the gene effect enrichment analysis results as previously described, there were significantly more disease groups enriched. As shown in Figure 3.10, for the paralog *NXT2* of the candidate gene *NXT1* there were 17 disease groups enriched.

3.3.6 Paralog mutation correlation

Correlations between deleterious mutations of a paralog and the dependency on the gene were calculated. In Figure 3.11 the distribution of correlation values for all candidate genes and their paralogs is shown. Most correlations lie around 0 indicating no strong correlation is present. To accommodate the fact that all correlations are generally close to zero, Z-scores are calculated based on the mean and standard deviation of the distribution shown in Figure 3.11.

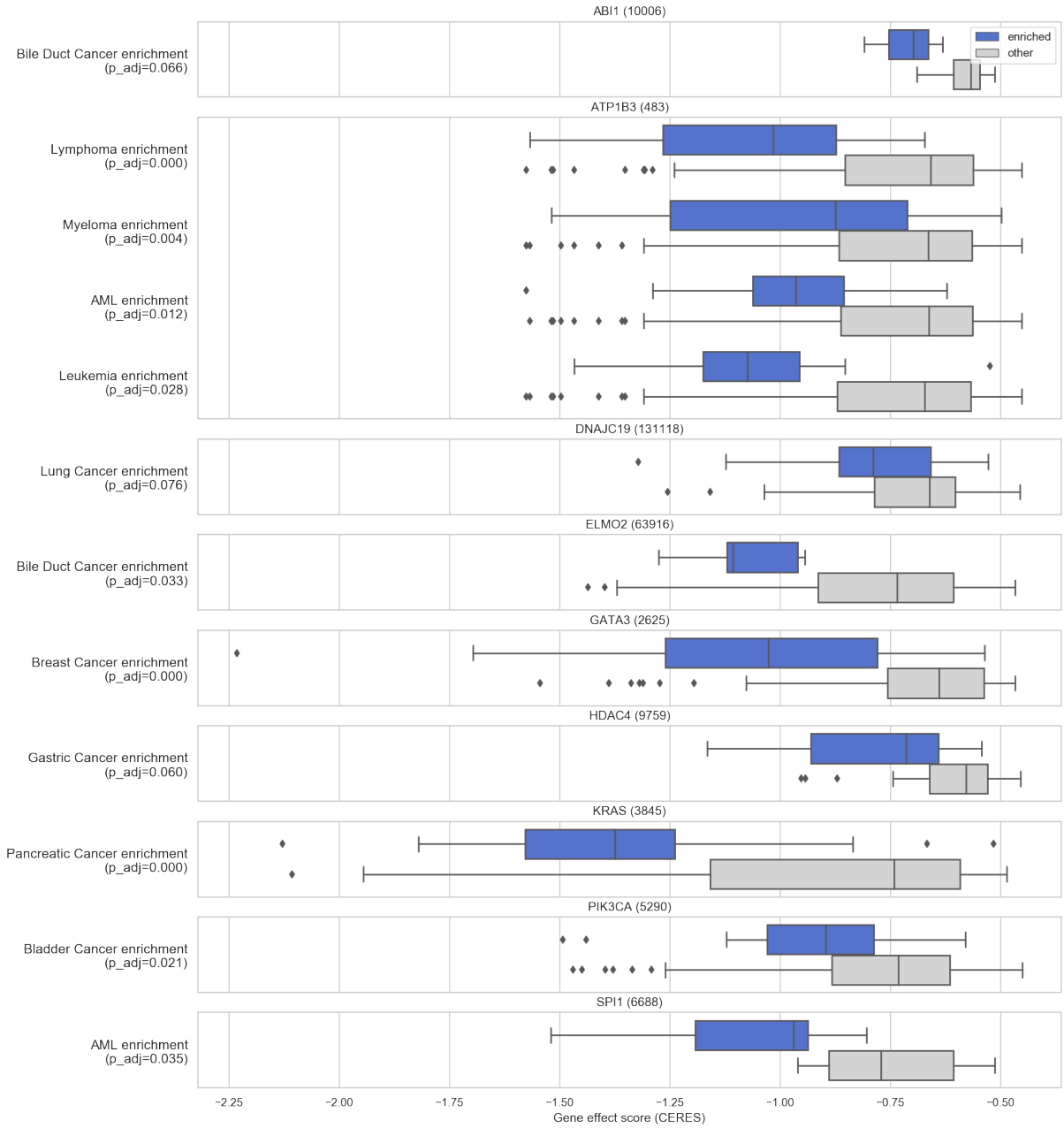


Figure 3.9: Enriched diseases for the final selection of candidate genes. Enrichment based on *gene effect* scores of cancer cell lines within specific disease groups for a candidate gene.

3.4 Attribute scoring

3.4.1 Significant diseases

Disease scores were calculated for each candidate gene from the introduced features. This score indicates the importance of the gene for that disease, i.e. a higher disease score means a higher relevance of the gene to the cancer type. For the final selection of candidates the disease scores are shown in Figure 3.12. For the final gene selection a disease score cutoff at the 95^{th} percentile was taken. This cutoff in the distribution of all candidate disease scores lay at 1.68.

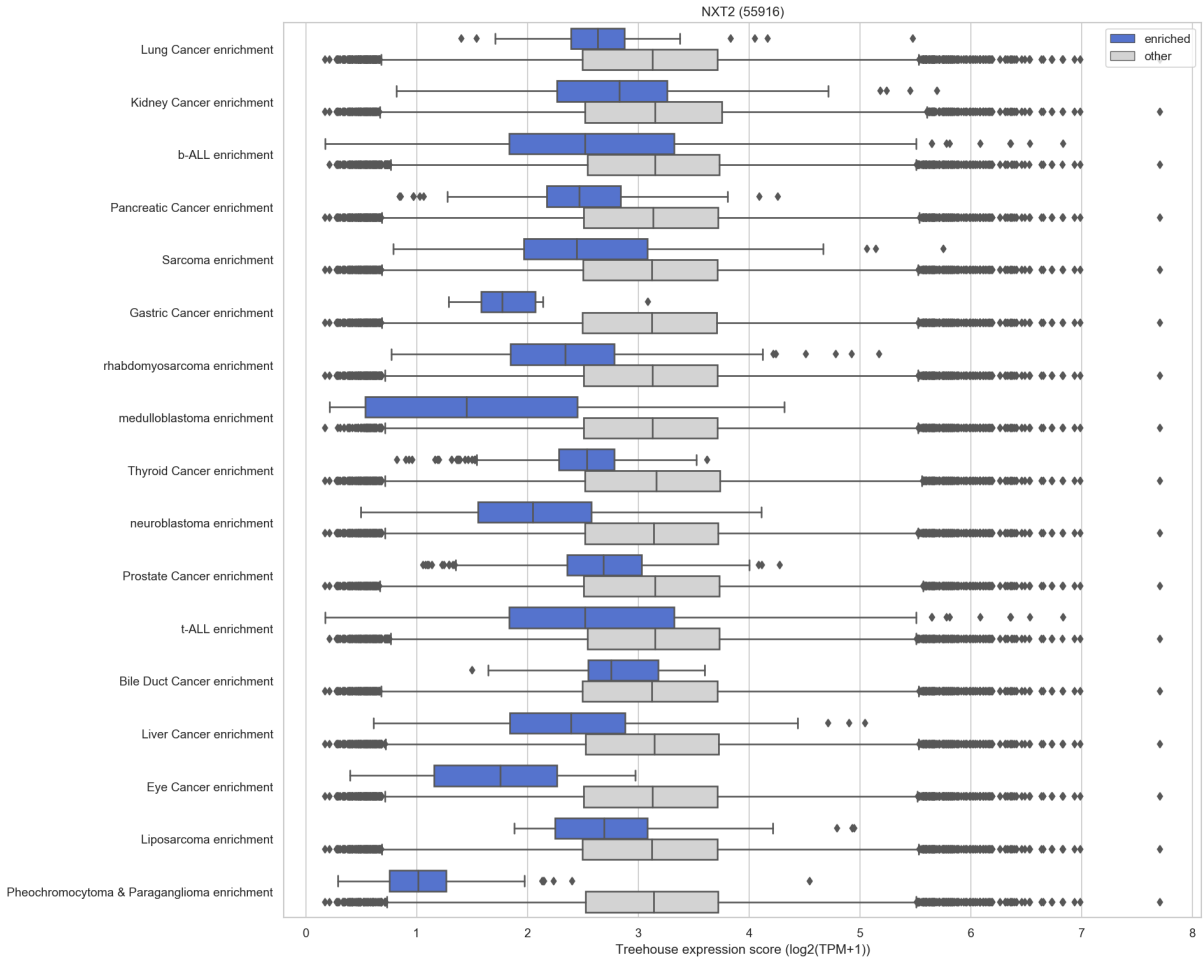


Figure 3.10: Enriched diseases for the paralog of *NXT1* based on the Treehouse [43] expression values. The Treehouse disease categories were mapped to DepMap disease names in the plot.

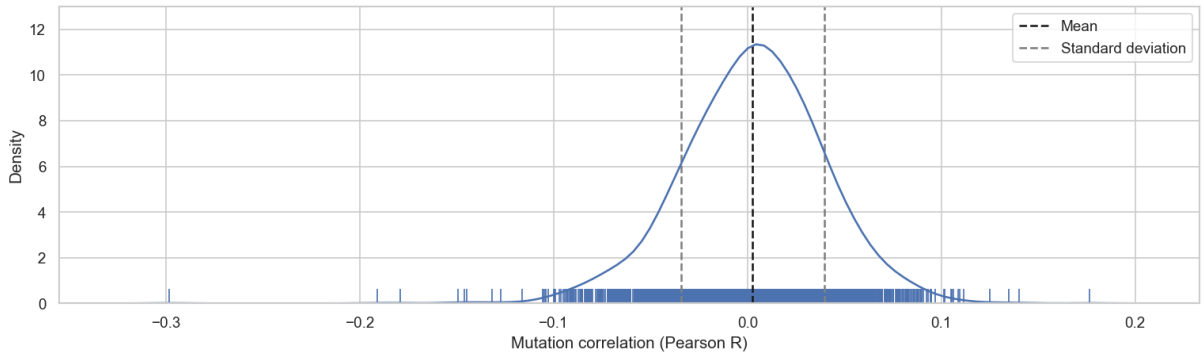


Figure 3.11: Distribution of gene effect to paralog mutation correlation values. A lower value indicates that the paralog is mutated more often in cancer cell lines where the candidate gene is more essential. The dashed vertical lines indicate the mean and standard deviations used to determine the z-scores.

3.4.2 Paralog scoring

Each paralog of the candidate genes was given a score to make a data supported discrimination between multiple paralogs and give a rough indication of which paralog is the most relevant. For each gene in the final selection of candidates its top paralog and the supporting features are given in Table 3.2.

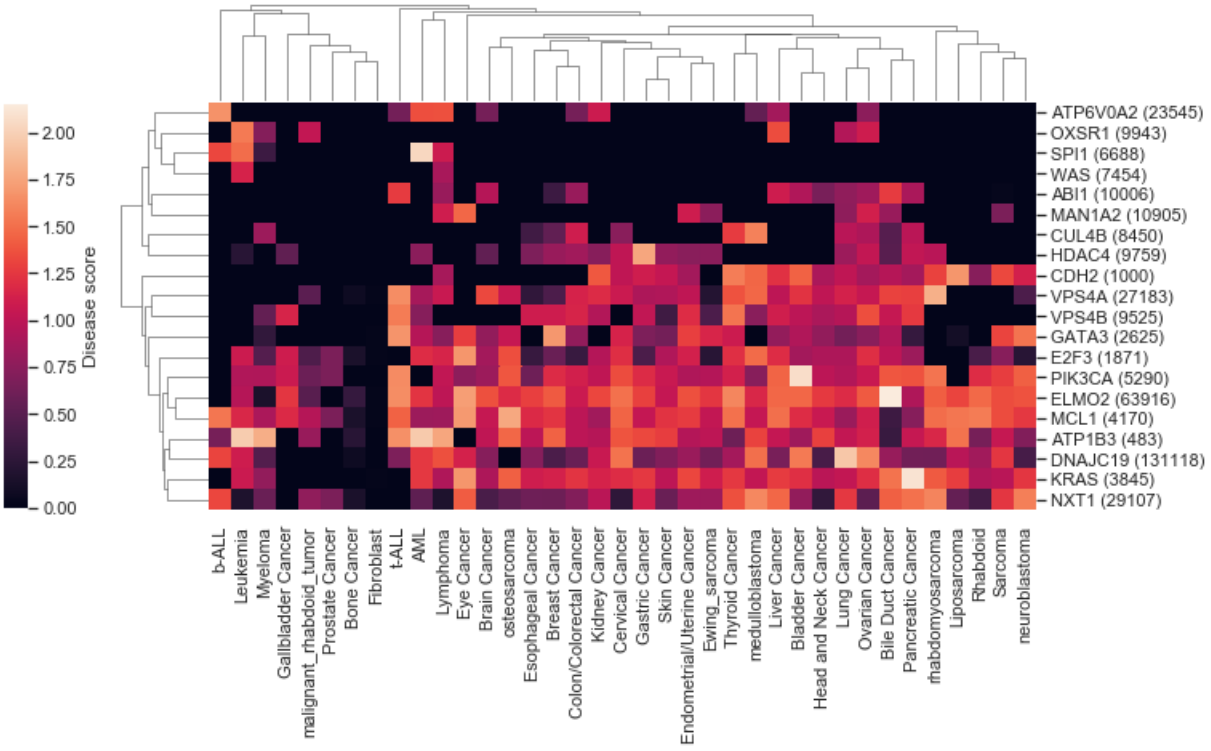


Figure 3.12: Disease scores per cancer type for the most interesting candidate genes. A higher score indicates a higher affinity of the gene for the cancer type.

Table 3.2: Overview of the top paralog for the final selection of most interesting candidate genes. A "–" indicates the paralog was not present in the DepMap best predicted genes. The mutation correlation column shows the Z-score of the correlation.

		Features					
Candidate gene	**Best paralog**	**Prediction**	**Mut. Cor.**	**Expr. Cor.**	**Max Dis. Cor.**	**Interacting**	**Score**
ABI1 (10006)	ABI2 (10152)	–	-0.31	0.25	0.63	yes	2.19
ATP1B3 (483)	ATP1B1 (481)	0.29	-0.20	0.54	0.70	yes	2.73
ATP6V0A2 (23545)	ATP6V0A1 (535)	0.58	-1.17	0.34	0.90	yes	3.99
CDH2 (1000)	DSC2 (1824)	–	-1.14	0.12	0.96	no	2.22
CUL4B (8450)	CUL4A (8451)	–	-2.01	0.25	1.00	yes	4.25
DNAJC19 (131118)	DNAJC15 (29103)	0.63	-1.43	0.59	0.91	yes	4.56
E2F3 (1871)	E2F4 (1874)	–	-0.40	0.13	0.75	yes	2.28
ELMO2 (63916)	ELMO1 (9844)	0.24	-1.01	0.24	0.95	yes	3.44
GATA3 (2625)	GATA4 (2626)	–	-1.01	-0.02	1.00	no	1.98
HDAC4 (9759)	HDAC10 (83933)	–	-2.11	-0.00	0.97	yes	4.07
KRAS (3845)	MRAS (22808)	–	-0.26	0.22	1.00	yes	2.48
MAN1A2 (10905)	MAN1C1 (57134)	–	-0.35	0.08	0.92	yes	2.34
MCL1 (4170)	BAX (581)	–	-1.92	0.04	0.99	yes	3.95
NXT1 (29107)	NXT2 (55916)	0.26	0.59	0.49	0.72	yes	1.88
OXSR1 (9943)	STK39 (27347)	0.71	-0.13	0.40	1.00	yes	3.23
PIK3CA (5290)	PIK3CD (5293)	0.19	-1.95	0.37	0.99	yes	4.49
SPI1 (6688)	ELF1 (1997)	–	-0.71	-0.15	0.98	yes	2.54
VPS4A (27183)	VPS4B (9525)	0.18	-0.28	0.32	0.91	yes	2.68
VPS4B (9525)	VPS4A (27183)	–	-2.11	0.34	0.97	yes	4.42
WAS (7454)	WASL (8976)	–	0.70	0.24	0.81	yes	1.35

3.5 Selection of genes of interest

From the 429 annotated candidate genes 20 genes of interest were selected based on the set feature cutoffs. The pipeline for filtering of candidates is shown in Figure 3.13. These most interesting candidates were already highlighted in previous results and a full list of their feature values is provided in the

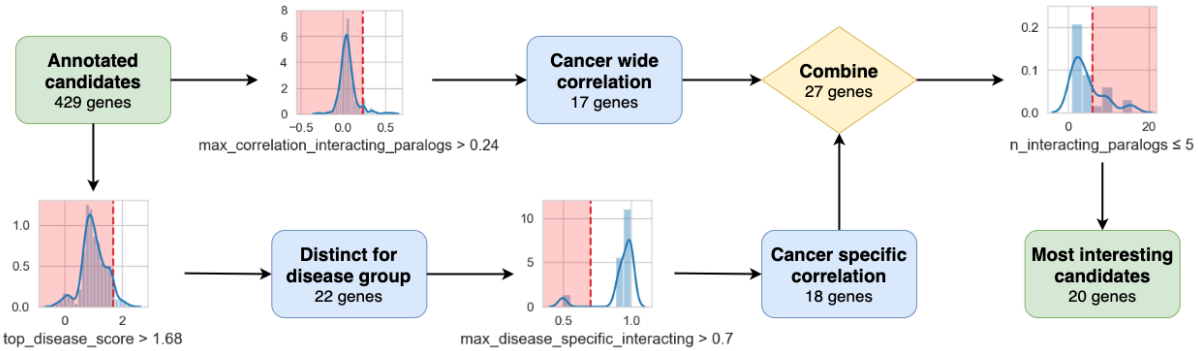


Figure 3.13: Diagram of the selection of most interesting candidate genes based on the introduced features.

Table 3.3: Final selection of most interesting candidate genes with several key features. Including, number of paralogs, number of interacting paralogs, number of common essentials the gene interacts with, number of cancer cell lines dependent on the gene, the maximal gene dependency interacting paralog expression correlation, top disease score, and the top scoring disease for the candidate gene.

Candidate gene	Paralogs	Int. Paralogs	Com. Ess.	Dep. Lines	Max Expr. Cor.	Top Dis. Score	Top Disease
ABI1 (10006)	2	2	16	17	0.25	1.27	Bile Duct Cancer
ATP1B3 (483)	3	3	3	281	0.54	1.97	Leukemia
ATP6V0A2 (23545)	2	2	23	18	0.34	1.67	b-ALL
CDH2 (1000)	22	4	8	59	0.25	1.68	Liposarcoma
CUL4B (8450)	5	5	77	19	0.25	1.59	medulloblastoma
DNAJC19 (131118)	1	1	19	144	0.59	1.94	Lung Cancer
E2F3 (1871)	5	5	12	192	0.13	1.69	Eye Cancer
ELMO2 (63916)	5	1	8	238	0.24	2.15	Bile Duct Cancer
GATA3 (2625)	5	2	20	107	0.10	1.69	Breast Cancer
HDAC4 (9759)	4	4	49	45	0.15	1.77	Gastric Cancer
KRAS (3845)	15	5	13	286	0.26	2.11	Pancreatic Cancer
MAN1A2 (10905)	3	2	11	15	0.24	1.47	Eye Cancer
MCL1 (4170)	6	3	3	421	0.05	1.78	osteosarcoma
NXT1 (29107)	1	1	47	267	0.49	1.61	medulloblastoma
OXSR1 (9943)	28	3	4	11	0.40	1.55	Leukemia
PIK3CA (5290)	5	5	22	273	0.37	2.07	Bladder Cancer
SPI1 (6688)	15	3	13	22	0.06	2.03	AML
VPS4A (27183)	4	1	16	159	0.32	1.82	rhabdomyosarcoma
VPS4B (9525)	4	1	18	93	0.34	1.56	t-ALL
WAS (7454)	3	2	13	3	0.24	1.13	Leukemia

supplementary materials, but an overview of their most key features is given in Table 3.3.

From the 20 final genes of interest, 14 genes were selected based on expression correlation: *ABI1*, *ATP1B3*, *ATP6V0A2*, *CDH2*, *CUL4B*, *DNAJC19*, *KRAS*, *MAN1A2*, *NXT1*, *OXSR1*, *PIK3CA*, *VPS4A*, *VPS4B* and *WAS*. Based on cancer specific essentiality only, 12 genes were selected: *ATP1B3*, *CDH2*, *DNAJC19*, *E2F3*, *ELMO2*, *GATA3*, *HDAC4*, *KRAS*, *MCL1*, *PIK3CA*, *SPI1* and *VPS4A*.

Chapter 4

Results: *NXT1* in Neuroblastoma

In this chapter the results of the in-depth study of *NXT1* in neuroblastoma are presented. It is investigated with what regions in the genome NXT1 interacts through ChIPseq analysis. Second, correspondence with the CRC in Kelly is investigated. Finally, the importance of *NXT1* interaction on gene expression is analyzed.

4.1 *NXT1* peak analysis

NXT1 ChIPseq data from the two treatments were compared, DMSO and dTAG. There were 5,576 peaks called in the DMSO Kelly sample. In the *NXT1* degraded sample significantly less interactions were found. Here 163 ChIPseq peaks were reported. In Figure 4.1 the correspondence between the two peak sets is shown.

The type of binding regions were retrieved trough peak annotation. The annotation of the NXT1 ChIPseq peaks is given in Figure 4.2. As shown in Figure 4.2a, NXT1 is located for the most part (about 70%) at promoter regions of genes. The few remaining regions where peaks are called after *NXT1* degradation are mostly at *distal intergenic* locations. The overlapping peaks in Figure 4.2c are a mix of mostly promoter regions and distal intergenic regions.

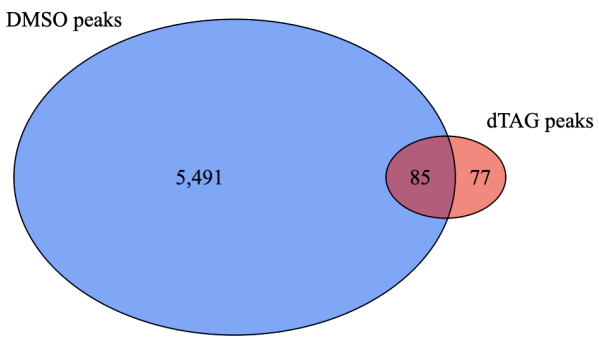
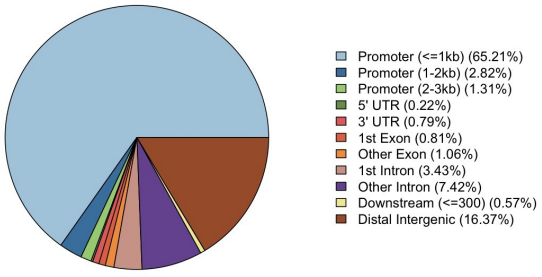
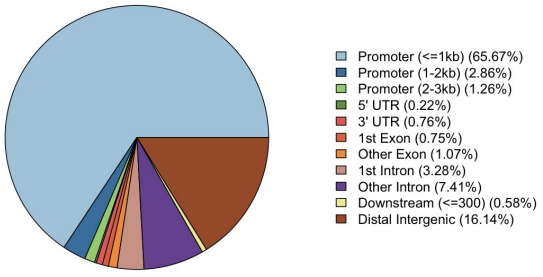


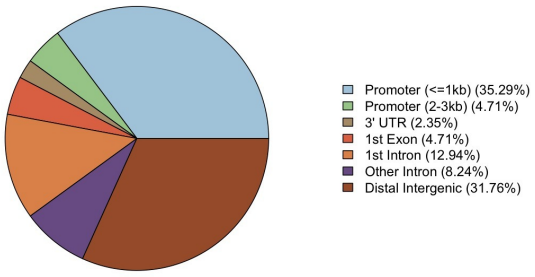
Figure 4.1: Comparison of ChIPseq peaks of NXT1 binding locations for the normal Kelly line and the NXT1 degraded system.



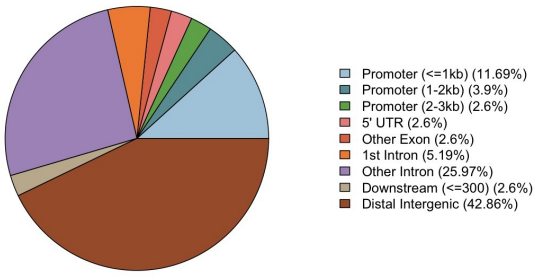
(a) All peaks



(b) DMSO only peaks



(c) Overlapping peaks



(d) dTAG only peaks

Figure 4.2: (a) Annotation of the type of binding locations of NXT1 in Kelly. The three annotations in (b), (c) and (d) correspond to the peak sets as separated in Figure 4.1.

Table 4.1: The number of overlapping peaks for each CRC member with NXT1. Also the number of acetylation (H3K27ac peaks) is given.

	MYCN	H3K27ac	TBX2	HAND2	GATA3	PHOX2B	ISL1
NXT1 (5,576 peaks)	4,882	4,797	2,154	1,879	1,488	972	691
Promoter peaks (3,866 peaks)	3,559	3,590	1,143	818	592	187	49
Other regions (1,710 peaks)	1,323	1,207	1,011	1,061	896	785	642

4.2 Overlap with CRC

CRC binding regions were compared to those of NXT1. The correspondence of peaks is shown in Figure 4.3 and quantitatively in Table 4.1. From the sorted ChIPseq binding regions in Figure 4.3 we see a large overlap of NXT1 promoter peaks with MYCN peaks. Most of the NXT1 promoter peaks are also binding regions of MYCN (Table 4.1). There is overlap of this category with other CRC members as well, but only very few (0.37%) of the NXT1 promoter peaks intersect with non-MYCN CRC peaks.

When comparing overlapping NXT1 peaks across all CRC members in Figure 4.4a, we see a second large intersection group of 589 peaks is NXT1 non-promoter peaks overlapping with *all* of the CRC members' peaks. This is the fourth (from the left) intersection group in Figure 4.4a. As shown in Figure 4.4c, these peaks are mostly at distal intergenic regions in the genome. A similar annotation is found for all

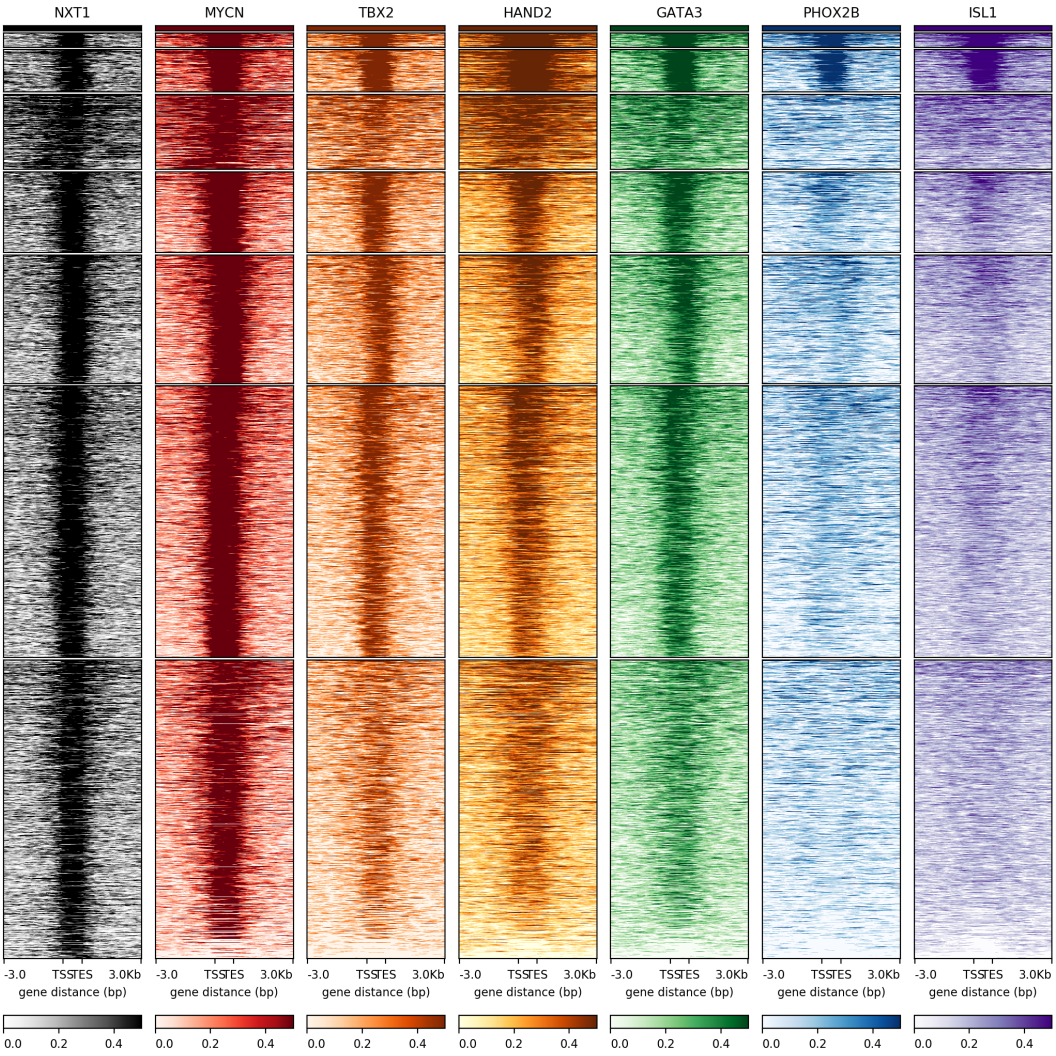


Figure 4.3: Correspondence of the aligned NXT1 ChIPseq reads to those of the CRC members. Regions (rows) are defined as binding locations of the protein in question and ranked by NXT1 signal in that region. Color keys for scaled reads-per-million-normalized signal are displayed below each heatmap. The rows were clustered using K-means clustering with 8 centers.

remaining NXT1 binding regions as is shown in Figure 4.4d.

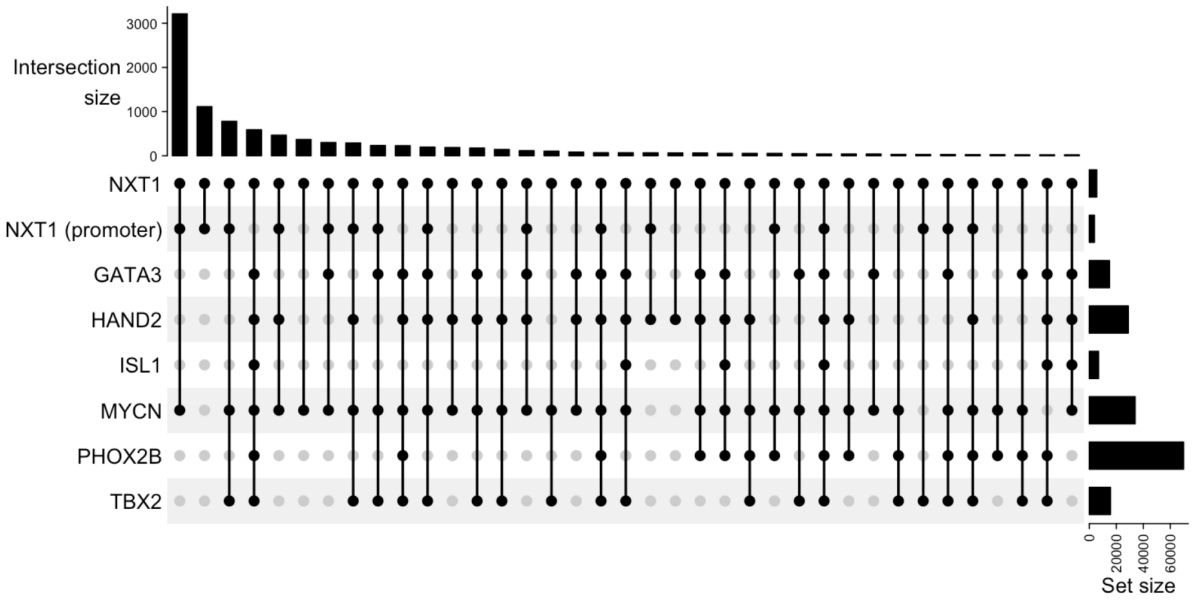
From the overlap analysis four NXT1 peak groups were defined which are visualized in Figure 4.4b. Looking at the largest intersection, we defined a first group of NXT1 peaks as NXT1–MYCN promoter peaks. The second peak group is defined by all remaining NXT1 promoter peaks. The next peak group consists of all NXT1 peaks overlapping with all of the CRC members. In the final group all remaining NXT1 peaks are stored. In subsequent these groups will be referred to as follows:

A. NXT1–MYCN promoter peaks (3,569 peaks),

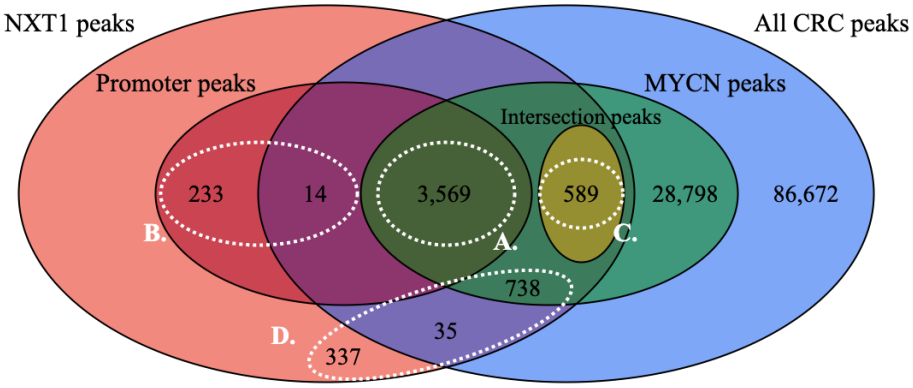
B. Other NXT1 promoter peaks (247 peaks),

C. NXT1–CRC peaks (589 peaks),

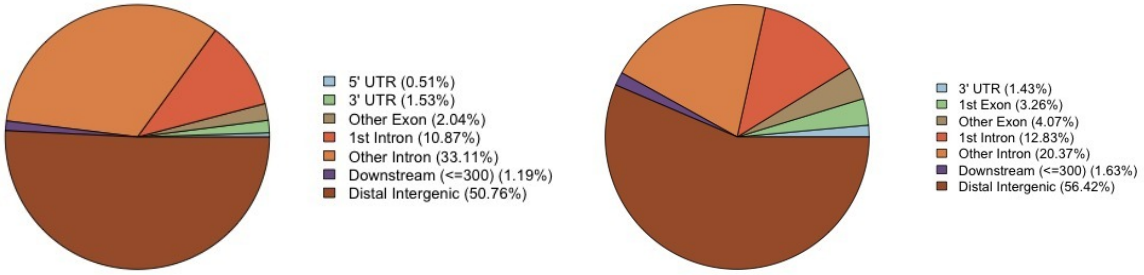
D. Other NXT1 peaks (1,110 peaks).



(a)

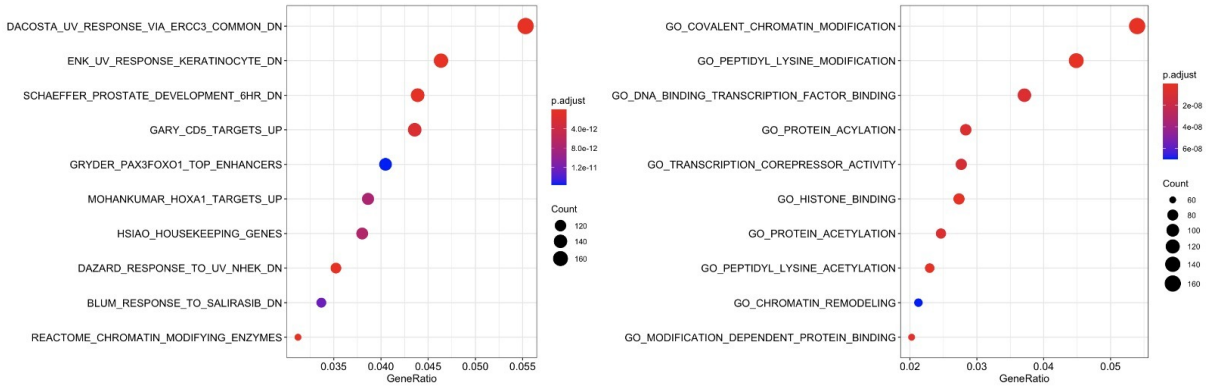


(b)

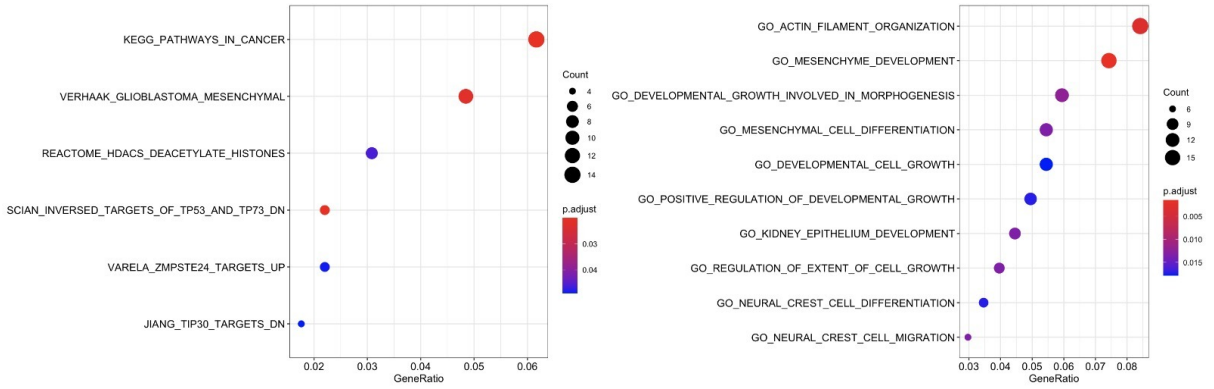


(c) (d)

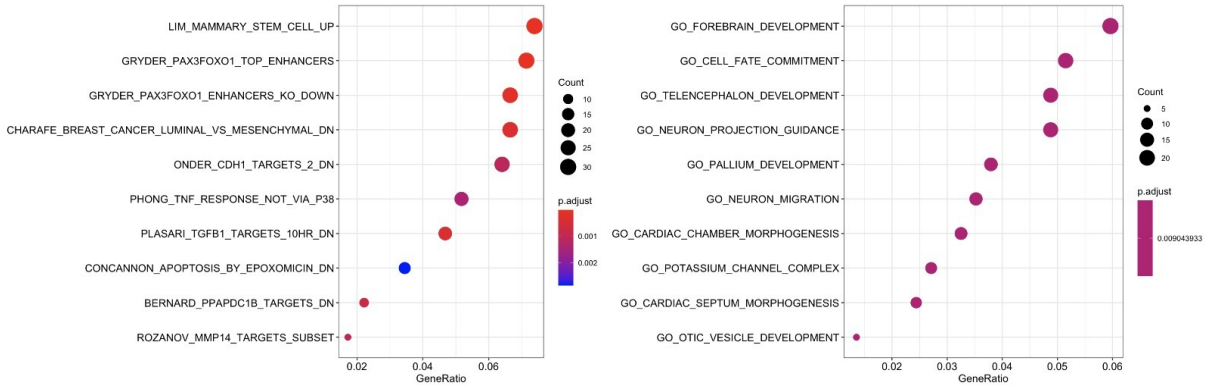
Figure 4.4: Overlap of NXT1 binding locations with those of the CRC members in Kelly. In (a) an upset plot shows all intersections with more than 20 peaks and sorted by size. Figure (b) shows an abstraction of the peak overlaps in a Venn diagram. Here the peak groups used in subsequent analysis are marked with white dotted lines. The yellow *intersection peaks* correspond to the fourth (from the left) intersection in (a). The annotation of these intersection peaks is given in (c). In (d) the genomic annotation of the peaks in group D are given.



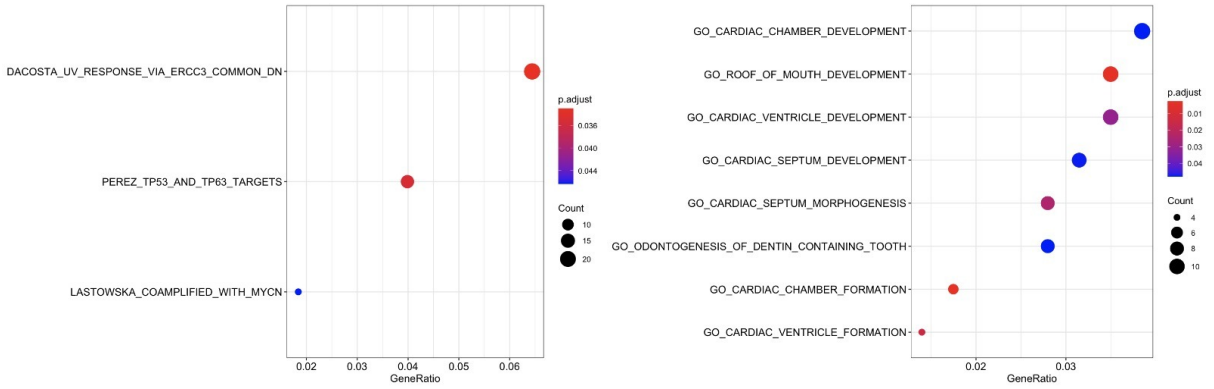
(a)



(b)



(c)



(d)

Figure 4.5: Gene sets enriched for the different NXT1 peak groups. Subfigures (a), (b), (c) and (d) correspond to the equally named peak groups. Figures on the left are enrichments in the MSigDB [75,76] curated gene sets, on the right are enrichments in the Gene Ontology [79,80] gene sets.

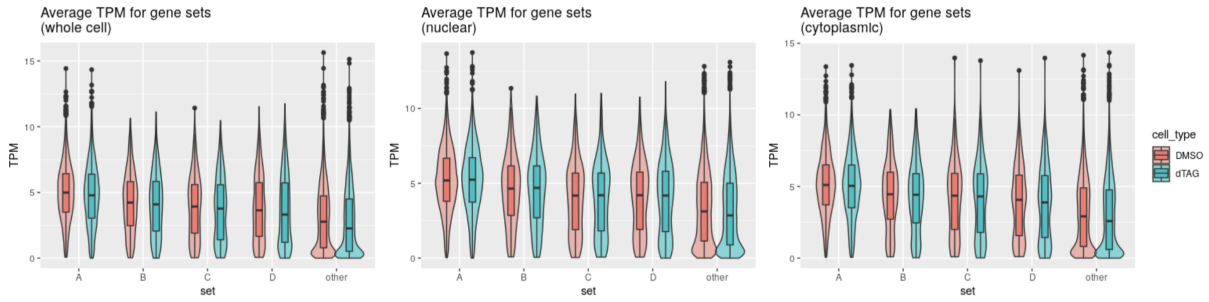


Figure 4.6: Comparison of RNAseq expression in average transcript per million (TPM) over the multiple cell groups in Kelly for the four different gene groups corresponding to the NXT1 binding locations.

4.2.1 Peak group GSEA

Gene set enrichment analysis (GSEA) was performed on the genes annotated to each peak in the peak groups to investigate their function. In Figure 4.5 the enrichments for the MSigDB [75,78] Curated and GO [79,80] gene sets are shown. Hallmark [78] enrichment analysis was also performed, but very few enrichments were found. The genes from peak groups C and D did not enrich for any Hallmark gene sets and group B enriched for only one gene set.

A rough distinction could be made between the enriched gene sets for peak groups mapping to promoter regions and those to other genomic regions. As Figure 4.5a and b show, the peak groups A and B were enriched mostly for more "general" cell cycle gene sets, such as several DNA repair (e.g. UV response) and housekeeping gene sets. For the exception of two enriched neural crest cell development gene sets in Figure 4.5b. The non-promoter peak groups C and D, on the other hand, enriched for more "neuroblastoma-like" gene sets. As illustrated by Figure 4.5c and d, many gene sets related to brain, cardiac and neuron development were enriched, which is where neuroblastoma mainly develops [3].

4.3 Gene expression effect

4.3.1 Baseline gene expression

A baseline of expression values per peak group was investigated by comparing their mean expression in TPM from all measurements. In Figure 4.6 we show that the average expression was higher for NXT1 bound genes, especially so for genes where NXT1 binds to the promoter region. Moreover, a decrease in expression was generally seen in the NXT1 degraded samples.

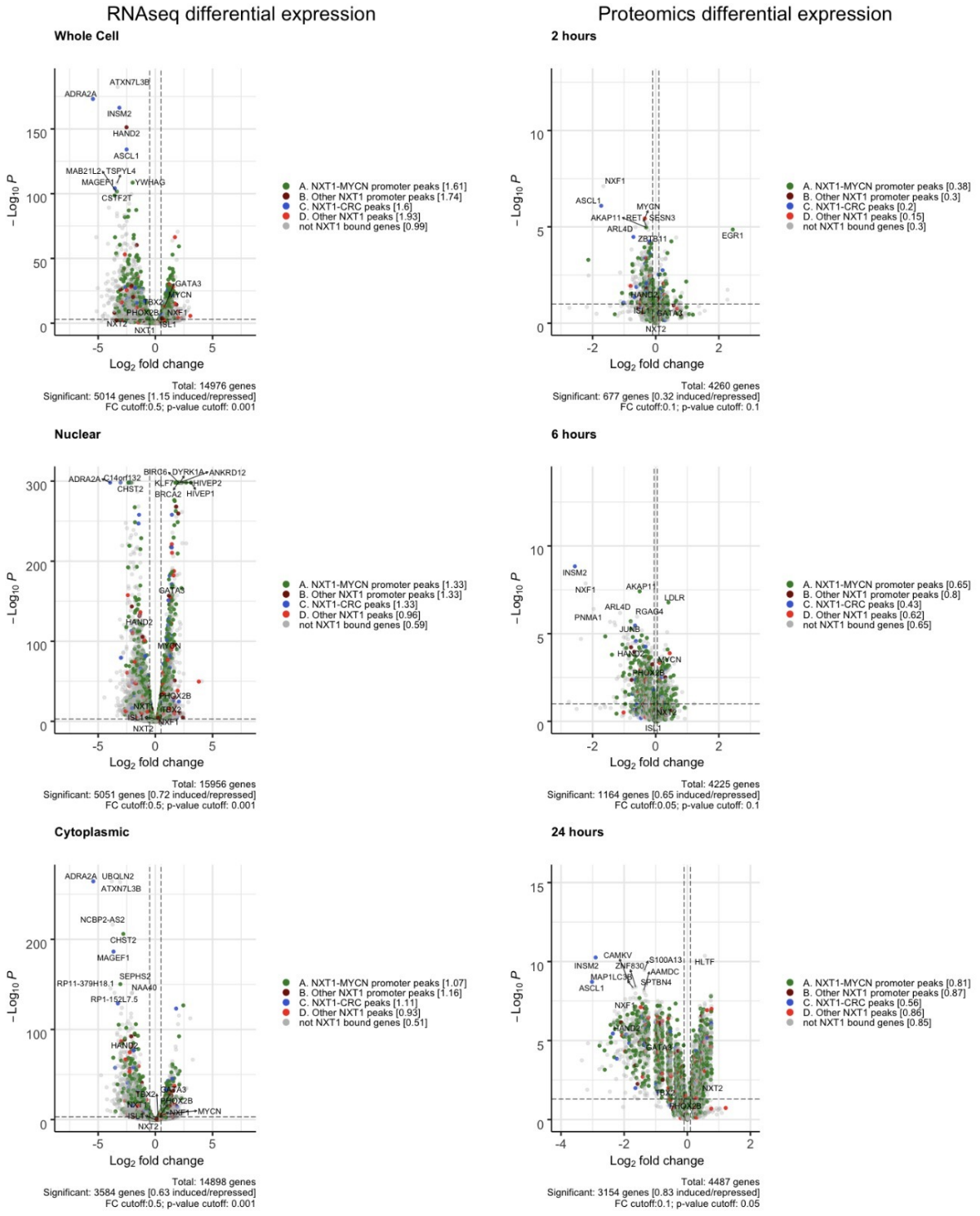
4.3.2 Differential expression

Next, change in expression was analyzed as visualized in Figure 4.7. From the RNAseq data we can see that the NXT1 bound genes were generally more induced than repressed after degradation. A major difference between the nuclear and cytoplasmic samples was in the scale of expression. The nuclear induced genes had generally much lower p-values (i.e. a higher $-log_{10}P$) compared to the cytoplasmic

genes. A lower p-value and higher fold change indicates that the gene had a greater expression change. Following this, the induced genes in the nuclear sample were clearly more heavily induced than those in the cytoplasmic sample, even though the ratios were relatively similar. Furthermore, the induced to repressed ratio of 0.63 indicated that there were almost twice as many repressed genes in the nuclear sample as compared to the whole cell.

When looking at the proteomics data, a different pattern can be seen, though. As is notably made clear by the 6 and 24 hour sample, the NXT1 bound genes were more repressed than induced. Here, especially peak group C showed a general repressed trend for the corresponding genes.

GSEA was performed on the groups of induced genes and repressed genes for both the whole cell RNAseq and 24 hours proteomics data. Here only the significant differentially expressed genes were used. The results of this can be found in appendix Figure C.1. There were no distinct differences found between the enriched gene sets for the repressed or induced genes.



(a) (b)

Figure 4.7: Differentially expression of genes after NXT1 degradation. The genes corresponding to NXT1 peak groups are colored and in the legend the number in square brackets indicates the ratio of number of induced genes to number of repressed genes. Subfigure (a) shows RNAseq expression in whole cell, nuclear and cytoplasmic. (b) is from MS proteomics data at three time points.

Chapter 5

Discussion

In this chapter all the results presented in chapters 3 and 4 are discussed. In addition, we discuss limitations with the current study, future directions and final conclusions.

5.1 Paralog genes interacting with common essentials

Paralog genes interacting with common essentials were identified by combining several state-of-the-art databases. After filtering on selective dependencies and annotating each gene with relevant features 429 candidate genes were found. There are several points of discussion on these results, however, which will be reviewed below.

5.1.1 STRING score

The STRING interaction score we defined based on a combination of the database, experiment and coexpression scores was validated against CORUM protein complexes. Although these complexes were treated as PPIs there are some caveats with this approach. One point of discussion is that protein complexes might not be represented in STRING in a sufficient manner. For instance, only 77% of all binary interactions from CORUM where described in STRING. Our defined STRING interaction score and cutoff leave have room for improvement.

The reasoning behind using only the database, experiment and coexpression scores was to reduce the number of false positive interactions in STRING. Subsequently, having a somewhat low cutoff of 0.6 would reduce the number of false negatives. However, the false positive and false negative rates were not strictly validated.

On the other hand, experiments with different STRING score cutoffs did not influence the final list of candidate genes significantly. This would suggest that the most important interactions generally have sufficiently high scores, or are also covered by one or more of the other PPI databases.

5.1.2 Paralog score

Although most (80%) of the candidate genes have ≤ 10 paralogs and ≤ 4 interacting paralogs, some candidates have over 50 paralogs. In order to make some differentiation between the paralogs based on relevance to essentiality in cancer, a method was desired to computationally identify the most *interesting* paralog. Our defined paralog score was a means of providing such an indication. We do not claim, however, that this score actually points to the most related paralog. In its current form the paralog score is a straight-forward sum of several features. But no weighting of any kind is applied yet. This would be essential for a more precise scoring method.

A good paralog score could help identify what we define as an *interaction triangle*: the interaction sub-network of a gene, its paralog and a common essential. This "simple" system is the ideal case for developing a targeting method as it highlights the essential molecular machine based on the common essential. In particular, treatment focusing on the gene would be effective in cells where the specific paralog is underexpressed.

5.1.3 Disease score

In order to establish a single metric representing the importance of a gene within a cancer type, the features relative dependency, DepMap gene effect enrichment, paralog expression correlation, and Treehouse paralog expression enrichment were combined into the disease score. Each of these features represented different aspects of the data and contributed to a score of the importance of a gene for a specific disease. In its current form the disease score gives a rough indication of what cancer type should be considered for further follow up investigation. The weighting of the various elements remains a heuristic, however, and optimization could improve the scoring.

5.1.4 Most interesting candidates

Finally, from the 429 candidate genes the 20 most interesting genes were selected. This was done based on two criteria: genes which are potentially good targets in a large range of cancer types (i.e. cancer wide), or genes which are selective for a certain disease in specific. The first was addressed based on expression correlation in all cancer cell lines. The second mainly based on top disease score. Only three of these 20 final candidates were present in both selections. These were the genes: *ATP1B3*, *CDH2* and *DNAJC19*. That they were present in both selections does not mean they are even more interesting than the rest, though. *CDH2*, for example, has a relatively low expression correlation across all cancer cell lines and is only just above the cutoff. Moreover, its highest disease score is for liposoma type in which it has only one dependent line. This only helps to show that manual inspection of all results remains a crucial step.

When looking at the max expression correlation alone, the top three scoring genes are *DNAJC19* (0.59), *ATP1B3* (0.54) and *NXT1* (0.49). These three genes have relatively high Pearson R scores for the expression of the paralog against the essentiality of the gene. This gives a strong indication that when

their paralog is underexpressed, there is no other compensation mechanism for their molecular system and they would hence be good therapeutic genetic targets. In the following we will give a brief overview of their function in the cell.

DNAJC19

DNAJC19 (also referred to as TIM14) interacts with the common essential PAM16 (also referred to as TIM16 or MAGMAS) and has one paralog *DNAJC15* (also referred to as *MCJ*). DNAJC19 is part of a big protein system for mitochondrial protein import from the cytoplasm. The core protein in this system is mtHsp70 which facilitates the transport. DNAJC19 stimulates the ATPase activity of mtHsp70 [82]. DNAJC19 forms a stable complex with PAM16 which is part of the overall transport system [83]. *PAM16* seems to play a key role in several cancers and their resistance to chemotherapy [84]. Moreover, it is shown that if *DNAJC19* is deactivated its function can be taken over by its paralog, *DNAJC15* [85,86]. Furthermore, there are several studies which indicate the importance of *PAM16* in cancer [87–90]

ATP1B3

The protein encoded by *ATP1B3* belongs to the family of ATPases beta chain proteins which are essential membrane proteins responsible for establishing and maintaining the electrochemical gradients of ions across the plasma membrane [91,92]. ATP1B3 interacts with three common essentials and has three paralogs (ATP1B4, ATP1B1 and ATP4B) which interact with two of the common essentials (ATP1A1 and APT2A2). *ATP1B3* has the highest expression correlation with its paralog *APT1B1* (Pearson R = 0.54), which would suggest that this gene is the most likely to buffer the molecular function of *ATP1B3* when it becomes unavailable. As shown in Figure 3.6a, the common essential ATP1A1 interacts with all three of the paralogs of ATP1B3 which would suggest that this is the most important interaction.

NXT1

The protein encoded by *NXT1* (also referred to as *p15*) interacts with 47 common essentials, but only *NXF1*, *KPNB1* and *NUP98* also interact with its paralog, *NXT2*. The strongest interaction is with *NXF1* (also referred to as *TAP*) on which there are several studies describing the interaction [23,93–95]. NXT1 forms a heterodimer with NXF1 to function as a nuclear RNA export system. This NXF1-NXT1 protein complex is highly associated with the TRanscription and EXport (TREX) multi-subunit complex essential for mRNA processing and export [96]. A recent study indicates these complexes have a transcript based specificity. NXF1-NXT1 is preferentially required for export of single- or few-exon transcripts with long exons or low GC-content. TREX is rather affects spliced transcripts with high GC-content [97].

It has been shown that the function of *NXT1* can be buffered by *NXT2* [98]. This is also shown by the strong correlation between *NXT1* dependency and *NXT2* expression of 0.49 (see Figure 3.7 and Table 3.3). *NXT1* could therefore serve as a good therapeutic target in cancer cell types where *NXT2* is naturally less present.

5.2 *NXT1* in neuroblastoma

The ChIPseq analysis of *NXT1* in the neuroblastoma line Kelly gives an indication at what genomic regions NXT1 is present. From the peak annotation results we see that there is a clear distinction between the binding locations of NXT1. The majority of the peaks are located at promoter regions and for these it goes that there is a large overlap with MYCN binding locations. In particular, 92% of the NXT1 peaks located at promoter regions are also MYCN peaks (see Table 4.1). This group of peaks enriches for relatively general gene sets, though.

It should be noted that the number of overlapping peaks in Table 4.1 and Figure 4.4b show slight discrepancies. Although the intersection sizes for both Figures were determined using ChIPpeakAnno [72], slightly different numbers of overlapping peaks were found when comparing two groups with each other as in Table 4.1 and Figure 4.4b and comparing four groups (NXT1 peaks, NXT1 promoter peaks, MYCN peaks and combined CRC peaks) with each other as in Figure 4.4b. This is caused by ChIPpeakAnno grouping peaks together when determining overlapping GRanges in R since a peak is a genomic range instead of a continuous point. For example, when comparing the sets $A = \{[1–2], [3–4], [5–6]\}$ and $B = \{[1–6]\}$ it returns it as one single point of intersection.

For the remainder of the NXT1 peaks, these lie mostly at distal intergenic regions. Distal intergenic regions can include binding to enhancers or super enhancers which is a key feature of the CRC as well [47] and is shown by the 70% peak overlap with H3K27ac. The concurrency with all of the CRC members in this case could also explain why the genes associated with this peak group are more neuroblastoma related than those related to the promoter regions where NXT1 binds.

This is emphasized even more when looking at the effect on the gene expression. Although the ratio's of number of induced to number of repressed genes are somewhat contradictory in the RNAseq and proteomics data, the indication that the genes associated with the NXT1/CRC non-promoter peaks are generally more repressed holds (see Figure 4.7b 24hr). Since the expression of these genes is apparently more dependent on the presence NXT1, this could explain why NXT1 plays an important role in neuroblastoma.

Chapter 6

Conclusion

In this work we presented a computational method to identify a novel group of genetic dependencies in cancer. Through combination of multiple databases, 429 paralog selective dependencies interacting with common essentials were selected. By creating advanced features and filtering on these, a group of 20 genes was selected for further investigation. In order to include genes selectively essential across all cancers and those with very cancer specific selectivity, we found that the best approach for gene selection was by filtering on either the expression correlation or our defined disease score.

Our proposed set of genes could serve as useful candidates for therapeutic genetic targeting as their essentiality is reliant on the presence and expression of their paralogs. Due to genetic changes in cancer causing genes to be particularly over- or under-expressed, the reliance on a specific other gene is powerful knowledge which can be leveraged for therapeutic targeting.

Several of the top-scoring candidate genes were studied in more detail. In particular, the behavior of the nuclear RNA export factor, *NXT1*, was investigated in more depth in the representative neuroblastoma cell line Kelly. ChIPseq data analysis revealed that the majority of binding of NXT1 occurred at promoter regions. For these NXT1 promoter peaks an 92% overlap was found with MYCN binding locations. For the remainder of the peaks, GSEA showed that the genes corresponding to the NXT1 binding locations which were also positions where all of the CRC members were bound, were enriched for many neuron development gene sets which would suggest distal intergenic binding of NXT1 is important for neuroblastoma development.

This latter group of genes also showed the most repressed protein expression when the effects of *NXT1* degradation on overall gene expression were studied. These findings suggest that *NXT1* plays a crucial role for the expression of neuroblastoma related genes and why therapeutic targeting of *NXT1* could be an effective treatment method for neuroblastoma.

Despite the fact that this in-depth study has been performed on only one of our identified genes, we can say that our method for identifying a novel class op potential therapeutic targets in cancer is effective.

A wide net has been cast on genetic dependencies but through a stringent filtering method we mark the most potent targets. Because of their interaction with a common essential, these genes are part of important biological systems for cancer development. Our disease score indicates the cancer types in which they would be interesting therapeutic targets, as their essentiality in these cell types is defined for a large part by the presence or expression of a paralog.

6.1 Future work

From the 429 a selection of 20 genes was made which are proposed for further evaluation in the lab. In particular, the genes *ATP1B3*, *DNAJC19* and *NXT1* are suggested as most interesting candidates for follow-up investigation. Recommended research includes studying the exact behavior of these genes in their top cancer types. For instance, it would be interesting to investigate the unique mechanics which cause their essentiality in the specific cancer types.

The list of 20 genes we reported as candidates for further investigation was based on cutoffs at expression correlation, disease score and number number of paralogs. The n-th percentiles were taken to extract the genes with most relevant scores. It should be noted that these percentile cutoffs are parameters for our selection pipeline and were set manually. Optimization of these parameters as well as addition of other features would be suggested as future work. Integration with other datasets could also improve the genes which are selected and optimize the false discovery rate whilst increasing the number of candidate genes.

Bibliography

[1] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

[2] M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz, "Discovery and saturation analysis of cancer genes across 21 tumour types," *Nature*, vol. 505, no. 7484, pp. 495–501, 2014.

[3] M. S. Irwin and J. R. Park, "Neuroblastoma: paradigm for precision medicine," *Pediatric Clinics*, vol. 62, no. 1, pp. 225–256, 2015.

[4] A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, *et al.*, "Defining a cancer dependency map," *Cell*, vol. 170, no. 3, pp. 564–576, 2017.

[5] R. M. Meyers, J. G. Bryan, J. M. McFarland, B. A. Weir, A. E. Sizemore, H. Xu, N. V. Dharia, P. G. Montgomery, G. S. Cowley, S. Pantel, *et al.*, "Computational correction of copy number effect improves specificity of crispr–cas9 essentiality screens in cancer cells," *Nature genetics*, vol. 49, no. 12, pp. 1779–1784, 2017.

[6] J. M. Dempster, J. Rossen, M. Kazachkova, J. Pan, G. Kugener, D. E. Root, and A. Tsherniak, "Extracting biological insights from the project achilles genome-scale crispr screens in cancer cell lines," *BioRxiv*, p. 720243, 2019.

[7] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, "A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity," *science*, vol. 337, no. 6096, pp. 816–821, 2012.

[8] J. Joung, S. Konermann, J. S. Gootenberg, O. O. Abudayyeh, R. J. Platt, M. D. Brigham, N. E. Sanjana, and F. Zhang, "Genome-scale crispr-cas9 knockout and transcriptional activation screening," *Nature protocols*, vol. 12, no. 4, pp. 828–863, 2017.

[9] C. Fellmann, B. G. Gowen, P.-C. Lin, J. A. Doudna, and J. E. Corn, "Cornerstones of crispr–cas in drug discovery and therapy," *Nature reviews Drug discovery*, vol. 16, no. 2, pp. 89–100, 2017.

[10] K. Shimada, J. L. Muhlich, and T. J. Mitchison, "A tool for browsing the cancer dependency map reveals functional connections between genes and helps predict the efficacy and selectivity of candidate cancer drugs," *bioRxiv*, 2019.

[11] E. V. Koonin, "Orthologs, paralogs, and evolutionary genomics," *Annu. Rev. Genet.*, vol. 39, pp. 309–338, 2005.

[12] B. De Kegel and C. J. Ryan, "Paralog buffering contributes to the variable essentiality of genes in cancer cell lines," *PLoS genetics*, vol. 15, no. 10, 2019.

[13] C. J. Lord, N. Quinn, and C. J. Ryan, "Integrative analysis of large-scale loss-of-function screens identifies robust cancer-associated genetic interactions," *Elife*, vol. 9, p. e58925, 2020.

[14] M. Cereda, T. P. Mourikis, and F. D. Ciccarelli, "Genetic redundancy, functional compensation, and cancer vulnerability," *Trends in Cancer*, vol. 2, no. 4, pp. 160–162, 2016.

[15] W. Sellers, "Functional genomics approaches to the discovery of paralog dependencies in cancer [abstract]," *Proceedings of the AACR-NCI-EORTC International Conference on Molecular Targets and Cancer Therapeutics*, vol. 18, no. 12, 2019.

[16] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, *et al.*, "Systematic functional analysis of the caenorhabditis elegans genome using rnai," *Nature*, vol. 421, no. 6920, pp. 231–237, 2003.

[17] Z. Gu, L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W.-H. Li, "Role of duplicate genes in genetic robustness against null mutations," *Nature*, vol. 421, no. 6918, pp. 63–66, 2003.

[18] B. Ewen-Campen, S. E. Mohr, Y. Hu, and N. Perrimon, "Accessing the phenotype gap: enabling systematic investigation of paralog functional complexity with crispr," *Developmental cell*, vol. 43, no. 1, pp. 6–9, 2017.

[19] S. R. Viswanathan, M. F. Nogueira, C. G. Buss, J. M. Krill-Burger, M. J. Wawer, E. Malolepsza, A. C. Berger, P. S. Choi, J. Shih, A. M. Taylor, *et al.*, "Genome-scale analysis identifies paralog lethality as a vulnerability of chromosome 1p loss in cancer," *Nature genetics*, vol. 50, no. 7, pp. 937–943, 2018.

[20] P. Van Der Lelij, S. Lieb, J. Jude, G. Wutz, C. P. Santos, K. Falkenberg, A. Schlattl, J. Ban, R. Schwentner, T. Hoffmann, *et al.*, "Synthetic lethality between the cohesin subunits stag1 and stag2 in diverse cancer contexts," *Elife*, vol. 6, p. e26980, 2017.

[21] L. Benedetti, M. Cereda, L. Monteverde, N. Desai, and F. D. Ciccarelli, "Synthetic lethal interaction between the tumour suppressor stag2 and its paralog stag1," *Oncotarget*, vol. 8, no. 23, p. 37619, 2017.

[22] H. Huang, B. Zhang, P. A. Hartenstein, J.-n. Chen, and S. Lin, "Nxt2 is required for embryonic heart development in zebrafish," *BMC developmental biology*, vol. 5, no. 1, p. 7, 2005.

[23] B. W. Guzik, L. Levesque, S. Prasad, Y.-C. Bor, B. E. Black, B. M. Paschal, D. Rekosh, and M.-L. Hammarskjöld, "Nxt1 (p15) is a crucial cellular cofactor in tap-dependent export of intron-containing rna in mammalian cells," *Molecular and cellular biology*, vol. 21, no. 7, pp. 2545–2554, 2001.

[24] H. L. Wiegand, G. A. Coburn, Y. Zeng, Y. Kang, H. P. Bogerd, and B. R. Cullen, "Formation of tap/nxt1 heterodimers activates tap-dependent nuclear mrna export by enhancing recruitment to nuclear pore complexes," *Molecular and cellular biology*, vol. 22, no. 1, pp. 245–256, 2002.

[25] S. Aibara, J. Katahira, E. Valkov, and M. Stewart, "The principal mrna nuclear export factor nxf1: Nxt1 forms a symmetric binding platform that facilitates export of retroviral cte-rna," *Nucleic acids research*, vol. 43, no. 3, pp. 1883–1893, 2015.

[26] C. F. Malone, N. V. Dharia, G. Kugener, B. Paolella, M. Rothberg, M. Abdusamad, A. Gonzalez, N. Dumont, S. Younger, D. Root, *et al.*, "Crispr-cas9 screens identify the nuclear export factor nxt1 as a novel therapeutic target in mycn-amplified neuroblastoma," 2019.

[27] M. Ghandi, F. W. Huang, J. Jané-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, *et al.*, "Next-generation characterization of the cancer cell line encyclopedia," *Nature*, vol. 569, no. 7757, pp. 503–508, 2019.

[28] C. Yu, A. M. Mannan, G. M. Yvone, K. N. Ross, Y.-L. Zhang, M. A. Marton, B. R. Taylor, A. Crenshaw, J. Z. Gould, P. Tamayo, *et al.*, "High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines," *Nature biotechnology*, vol. 34, no. 4, p. 419, 2016.

[29] Broad-DepMap, "DepMap 20Q2 Public," 6 2020.

[30] J. M. McFarland, Z. V. Ho, G. Kugener, J. M. Dempster, P. G. Montgomery, J. G. Bryan, J. M. Krill-Burger, T. M. Green, F. Vazquez, J. S. Boehm, *et al.*, "Improved estimation of cancer dependencies from large-scale rnai screens using model-based normalization and data integration," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[31] F. M. Behan, F. Iorio, G. Picco, E. Gonçalves, C. M. Beaver, G. Migliardi, R. Santos, Y. Rao, F. Sassi, M. Pinnelli, *et al.*, "Prioritization of cancer therapeutic targets using crispr–cas9 screens," *Nature*, vol. 568, no. 7753, p. 511, 2019.

[32] Broad-DepMap, "Project SCORE processed with CERES," 8 2019.

[33] E. R. McDonald III, A. De Weck, M. R. Schlabach, E. Billy, K. J. Mavrakis, G. R. Hoffman, D. Belur, D. Castelletti, E. Frias, K. Gampa, *et al.*, "Project drive: a compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep rnai screening," *Cell*, vol. 170, no. 3, pp. 577–592, 2017.

[34] R. Marcotte, A. Sayad, K. R. Brown, F. Sanchez-Garcia, J. Reimand, M. Haider, C. Virtanen, J. E. Bradner, G. D. Bader, G. B. Mills, *et al.*, "Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance," *Cell*, vol. 164, no. 1-2, pp. 293–309, 2016.

[35] B. Li and C. N. Dewey, "Rsem: accurate transcript quantification from rna-seq data with or without a reference genome," *BMC bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[36] J. M. Dempster, J. Krill-Burger, A. Warren, J. McFarland, T. Golub, and A. Tsherniak, "Gene expression has more power for predicting in vitro cancer cell vulnerabilities than genomics," *bioRxiv*, 2020.

[37] M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and A. Ruepp, "Corum: the comprehensive resource of mammalian protein complexes—2019," *Nucleic acids research*, vol. 47, no. D1, pp. D559–D563, 2019.

[38] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, *et al.*, "String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research*, vol. 47, no. D1, pp. D607–D613, 2019.

[39] L. Licata, P. Lo Surdo, M. Iannuccelli, A. Palma, E. Micarelli, L. Perfetto, D. Peluso, A. Calderone, L. Castagnoli, and G. Cesareni, "Signor 2.0, the signaling network open resource 2.0: 2019 update," *Nucleic acids research*, vol. 48, no. D1, pp. D504–D510, 2020.

[40] K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, *et al.*, "A reference map of the human binary protein interactome," *Nature*, vol. 580, no. 7803, pp. 402–408, 2020.

[41] H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas, "Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools," *Nucleic acids research*, vol. 47, no. D1, pp. D419–D426, 2019.

[42] M. Ouedraogo, C. Bettembourg, A. Bretaudeau, O. Sallou, C. Diot, O. Demeure, and F. Lecerf, "The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes," *PloS one*, vol. 7, no. 11, 2012.

[43] "Treehouse childhood cancer initiative at the uc santa cruz genomics institute." `https://treehousegenomics.soe.ucsc.edu/`. Accessed: 06-29-2020.

[44] B. Nabet, J. M. Roberts, D. L. Buckley, J. Paulk, S. Dastjerdi, A. Yang, A. L. Leggett, M. A. Erb, M. A. Lawlor, A. Souza, *et al.*, "The dtag system for immediate and target-specific protein degradation," *Nature chemical biology*, vol. 14, no. 5, pp. 431–441, 2018.

[45] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-dna interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[46] A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow, "Genome-wide analysis of transcription factor binding sites based on chip-seq data," *Nature methods*, vol. 5, no. 9, pp. 829–834, 2008.

[47] A. D. Durbin, M. W. Zimmerman, N. V. Dharia, B. J. Abraham, A. B. Iniguez, N. Weichert-Leahey, S. He, J. M. Krill-Burger, D. E. Root, F. Vazquez, *et al.*, "Selective gene dependencies in mycn-amplified neuroblastoma include the core transcriptional regulatory circuitry," *Nature genetics*, vol. 50, no. 9, pp. 1240–1246, 2018.

[48] V. Saint-André, A. J. Federation, C. Y. Lin, B. J. Abraham, J. Reddy, T. I. Lee, J. E. Bradner, and R. A. Young, "Models of human core transcriptional regulatory circuitries," *Genome research*, vol. 26, no. 3, pp. 385–396, 2016.

[49] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, *et al.*, "Core transcriptional regulatory circuitry in human embryonic stem cells," *cell*, vol. 122, no. 6, pp. 947–956, 2005.

[50] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, *et al.*, "Histone h3k27ac separates active from poised enhancers and predicts developmental state," *Proceedings of the National Academy of Sciences*, vol. 107, no. 50, pp. 21931–21936, 2010.

[51] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014.

[52] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[53] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, *et al.*, "Jupyter notebooks-a publishing format for reproducible computational workflows.," in *ELPUB*, pp. 87–90, 2016.

[54] The pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020.

[55] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference* (Stéfan van der Walt and Jarrod Millman, eds.), pp. 56 – 61, 2010.

[56] A. D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, *et al.*, "Ensembl 2020," *Nucleic acids research*, vol. 48, no. D1, pp. D682–D688, 2020.

[57] U. Consortium, "Uniprot: a worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.

[58] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, *et al.*, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D733–D745, 2016.

[59] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at ncbi," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D52–D57, 2010.

[60] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, *et al.*, "The consensus coding sequence (ccds) project: Identifying a common protein-coding gene set for the human and mouse genomes," *Genome research*, vol. 19, no. 7, pp. 1316–1323, 2009.

[61] A. Alibés, P. Yankilevich, R. Díaz-Uriarte, *et al.*, "Idconverter and idclight: conversion and annotation of gene and protein ids," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–9, 2007.

[62] F. Al-Shahrour, P. Minguez, J. Tárraga, D. Montaner, E. Alloza, J. M. Vaquerizas, L. Conde, C. Blaschke, J. Vera, and J. Dopazo, "Babelomics: a systems biology perspective in the functional annotation of genome-scale experiments," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W472–W476, 2006.

[63] K. J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, J. N. Weinstein, *et al.*, "Matchminer: a tool for batch navigation among gene and gene product identifiers," *Genome biology*, vol. 4, no. 4, p. R27, 2003.

[64] B. T. S. Da Wei Huang, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "David gene id conversion tool," *Bioinformation*, vol. 2, no. 10, p. 428, 2008.

[65] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo, "g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments," *Nucleic acids research*, vol. 35, no. suppl_2, pp. W193–W200, 2007.

[66] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[67] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[68] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[69] T. E. Oliphant, *A guide to NumPy*, vol. 1. Trelgol Publishing USA, 2006.

[70] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[71] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, *et al.*, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019.

[72] L. J. Zhu, C. Gazin, N. D. Lawson, H. Pagès, S. M. Lin, D. S. Lapointe, and M. R. Green, "Chippeakanno: a bioconductor package to annotate chip-seq and chip-chip data," *BMC bioinformatics*, vol. 11, no. 1, p. 237, 2010.

[73] Z. Gu, R. Eils, and M. Schlesner, "Complex heatmaps reveal patterns and correlations in multidimensional genomic data," *Bioinformatics*, vol. 32, no. 18, pp. 2847–2849, 2016.

[74] G. Yu, L.-G. Wang, and Q.-Y. He, "Chipseeker: an r/bioconductor package for chip peak annotation, comparison and visualization," *Bioinformatics*, vol. 31, no. 14, pp. 2382–2383, 2015.

[75] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.

[76] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (msigdb) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[77] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an r package for comparing biological themes among gene clusters," *Omics: a journal of integrative biology*, vol. 16, no. 5, pp. 284–287, 2012.

[78] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The molecular signatures database hallmark gene set collection," *Cell systems*, vol. 1, no. 6, pp. 417–425, 2015.

[79] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[80] Gene Ontology Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic acids research*, vol. 47, no. D1, pp. D330–D338, 2019.

[81] K. Blighe, S. Rana, and M. Lewis, "Enhancedvolcano: Publication-ready volcano plots with enhanced colouring and labeling," *R package version*, vol. 1, no. 0, 2019.

[82] D. Mokranjac, M. Sichting, D. Popov-Čeleketič, A. Berg, K. Hell, and W. Neupert, "The import motor of the yeast mitochondrial tim23 preprotein translocase contains two different j proteins, tim14 and mdj2," *Journal of Biological Chemistry*, vol. 280, no. 36, pp. 31608–31614, 2005.

[83] S. Elsner, D. Simian, O. Iosefson, M. Marom, and A. Azem, "The mitochondrial protein translocation motor: structural conservation between the human and yeast tim14/pam18-tim16/pam16 co-chaperones," *International Journal of Molecular Sciences*, vol. 10, no. 5, pp. 2041–2053, 2009.

[84] E. Riva, F. Tagliati, T. Gagliano, C. Trapella, S. Missiroli, P. Pinton, C. Bertolucci, Z. M. Chiara, *et al.*, "Magmas modulates chemoresistance in endocrine-related cancers," in *19th European Congress of Endocrinology*, vol. 49, BioScientifica, 2017.

[85] C. Schusdziarra, M. Blamowska, A. Azem, and K. Hell, "Methylation-controlled j-protein mcj acts in the import of proteins into human mitochondria," *Human molecular genetics*, vol. 22, no. 7, pp. 1348–1357, 2013.

[86] V. Shridhar, K. C. Bible, J. Staub, R. Avula, Y. K. Lee, K. Kalli, H. Huang, L. C. Hartmann, S. H. Kaufmann, and D. I. Smith, "Loss of expression of a new member of the dnaj protein family confers resistance to chemotherapeutic agents used in the treatment of ovarian cancer," *Cancer research*, vol. 61, no. 10, pp. 4258–4265, 2001.

[87] P. T. Jubinsky, A. Messer, J. Bender, R. E. Morris, G. M. Ciraolo, D. P. Witte, R. G. Hawley, and M. K. Short, "Identification and characterization of magmas, a novel mitochondria-associated protein involved in granulocyte-macrophage colony-stimulating factor signal transduction," *Experimental hematology*, vol. 29, no. 12, pp. 1392–1402, 2001.

[88] P. T. Jubinsky, M. K. Short, G. Mutema, R. E. Morris, G. M. Ciraolo, and M. Li, "Magmas expression in neoplastic human prostate," *Journal of molecular histology*, vol. 36, no. 1-2, pp. 69–75, 2005.

[89] P. T. Jubinsky, M. K. Short, G. Mutema, and D. P. Witte, "Developmental expression of magmas in murine tissues and its co-expression with the gm-csf receptor," *Journal of Histochemistry & Cytochemistry*, vol. 51, no. 5, pp. 585–596, 2003.

[90] J. Peng, C.-H. Huang, M. K. Short, and P. T. Jubinsky, "Magmas gene structure and evolution," *In silico biology*, vol. 5, no. 3, pp. 251–263, 2005.

[91] N. Malik, V. A. Canfield, M.-C. Beckers, P. Gros, and R. Levenson, "Identification of the mammalian na, k-atpase $\beta 3$ subunit," *Journal of Biological Chemistry*, vol. 271, no. 37, pp. 22754–22758, 1996.

[92] N. Malik, V. Canfield, G. Sanchez-Watts, A. G. Watts, S. Scherer, B. G. Beatty, P. Gros, and R. Levenson, "Structural organization and chromosomal localization of the human na, k-atpase $\beta 3$ subunit gene and pseudogene," *Mammalian genome*, vol. 9, no. 2, p. 136, 1998.

[93] B. E. Black, L. Lévesque, J. M. Holaska, T. C. Wood, and B. M. Paschal, "Identification of an ntf2-related factor that binds ran-gtp and regulates nuclear protein export," *Molecular and cellular biology*, vol. 19, no. 12, pp. 8616–8624, 1999.

[94] B. Ossareh-Nazari, C. Maison, B. E. Black, L. Lévesque, B. M. Paschal, and C. Dargemont, "Rangtp-binding protein nxt1 facilitates nuclear export of different classes of rna in vitro," *Molecular and Cellular Biology*, vol. 20, no. 13, pp. 4562–4571, 2000.

[95] J. Katahira, H. Inoue, E. Hurt, and Y. Yoneda, "Adaptor aly and co-adaptor thoc5 function in the tap-p15-mediated nuclear export of hsp70 mrna," *The EMBO journal*, vol. 28, no. 5, pp. 556–567, 2009.

[96] C. G. Heath, N. Viphakone, and S. A. Wilson, "The role of trex in gene expression and disease," *Biochemical Journal*, vol. 473, no. 19, pp. 2911–2935, 2016.

[97] B. Zuckerman, M. Ron, M. Mikl, E. Segal, and I. Ulitsky, "Gene architecture and sequence composition underpin selective dependency of nuclear export of long rnas on nxf1 and the trex complex," *Molecular Cell*, 2020.

[98] C. F. Malone, N. F. Dharia, G. Kugener, M. V. Rothberg, M. Abdusumad, A. Gonzalez, M. Kuljanin, A. L. Robichaud, A. Saur Conway, J. M. Dempster, B. R. Paolella, N. Dumont, J. D. Mancias, S. T. Younger, D. E. Root, T. R. Golub, F. Vazquez, and K. Stegmaier, "Crispr screens identify selective modulation of a pan-essential protein as a therapeutic strategy in cancer," *under review*, 2020.

Appendix A

List of cancer types

Table A.1: List of our 43 defined disease types with the number of cell lines available in the DepMap file `sample_info.csv`

Disease name	Cancer cell lines
Adrenal Cancer	1
AML	53
b-ALL	27
Bile Duct Cancer	36
Bladder Cancer	39
Bone Cancer	13
Brain Cancer	95
Breast Cancer	82
Cervical Cancer	22
Colon/Colorectal Cancer	83
Embryonal Cancer	3
Endometrial/Uterine Cancer	39
Engineered	10
Esophageal Cancer	38
Ewing_sarcoma	49
Eye Cancer	9
Fibroblast	43
Gallbladder Cancer	7
Gastric Cancer	49
Head and Neck Cancer	76
Kidney Cancer	55
Leukemia	29

Liposarcoma	11
Liver Cancer	27
Lung Cancer	273
Lymphoma	109
malignant_rhabdoid_tumor	16
medulloblastoma	12
Myeloma	34
neuroblastoma	46
Non-Cancerous	5
osteosarcoma	17
Ovarian Cancer	74
Pancreatic Cancer	59
Prostate Cancer	13
Rhabdoid	5
rhabdomyosarcoma	19
Sarcoma	23
Skin Cancer	113
t-ALL	23
Teratoma	1
Thyroid Cancer	21
Unknown	45

Appendix B

Treehouse to DepMap disease name mapping

Table B.1: The Treehouse disease names as translated to our, DepMap based, disease types.

Treehouse disease	DepMap disease
acute leukemia	Leukemia
acute lymphoblastic leukemia	b-ALL
acute lymphoblastic leukemia	t-ALL
acute megakaryoblastic leukemia	AML
acute myeloid leukemia	AML
adrenocortical carcinoma	Adrenal Cancer
alveolar rhabdomyosarcoma	rhabdomyosarcoma
bladder urothelial carcinoma	Bladder Cancer
breast invasive carcinoma	Breast Cancer
cervical & endocervical cancer	Cervical Cancer
cholangiocarcinoma	Bile Duct Cancer
choroid plexus carcinoma	Brain Cancer
colon adenocarcinoma	Colon/Colorectal Cancer
dedifferentiated liposarcoma	Liposarcoma
desmoplastic small round cell tumor	Sarcoma
diffuse large B-cell lymphoma	Lymphoma
dysembryoplastic neuroepithelial tumor	Brain Cancer
embryonal rhabdomyosarcoma	rhabdomyosarcoma
ependymoma	Brain Cancer
esophageal carcinoma	Esophageal Cancer

Ewing sarcoma	Ewing_sarcoma
fibrolamellar hepatocellular carcinoma	Liver Cancer
gastrointestinal stromal tumor	Gastric Cancer
glioblastoma multiforme	Brain Cancer
glioma	Brain Cancer
head & neck squamous cell carcinoma	Head and Neck Cancer
hepatoblastoma	Liver Cancer
hepatocellular carcinoma	Liver Cancer
kidney chromophobe	Kidney Cancer
kidney clear cell carcinoma	Kidney Cancer
kidney papillary cell carcinoma	Kidney Cancer
leiomyosarcoma	Sarcoma
lung adenocarcinoma	Lung Cancer
lung squamous cell carcinoma	Lung Cancer
lymphoma	Lymphoma
malignant peripheral nerve sheath tumor	Malignant peripheral nerve sheath tumor
medulloblastoma	medulloblastoma
melanoma	Skin Cancer
mesothelioma	Lung Cancer
myxofibrosarcoma	Sarcoma
neuroblastoma	neuroblastoma
osteosarcoma	osteosarcoma
ovarian serous cystadenocarcinoma	Ovarian Cancer
pancreatic adenocarcinoma	Pancreatic Cancer
pheochromocytoma & paraganglioma & Pheochromocytoma	Paraganglioma
prostate adenocarcinoma	Prostate Cancer
rectum adenocarcinoma	Colon/Colorectal Cancer
rhabdomyosarcoma	rhabdomyosarcoma
sarcoma	Sarcoma
skin cutaneous melanoma	Skin Cancer
stomach adenocarcinoma	Gastric Cancer
supratentorial embryonal tumor NOS	Brain Cancer
synovial sarcoma	Synovial sarcoma
testicular germ cell tumor	Germ Cell Tumor
thymoma	Thymoma
thyroid carcinoma	Thyroid Cancer
undifferentiated pleomorphic sarcoma	Sarcoma

undifferentiated sarcoma NOS	Sarcoma
uterine carcinosarcoma	Endometrial/Uterine Cancer
uterine corpus endometrioid carcinoma	Endometrial/Uterine Cancer
uveal melanoma	Eye Cancer
wilms tumor	Kidney Cancer
acute leukemia of ambiguous lineage	Leukemia
myoepithelial carcinoma	Myoepithelial Carcinoma
INI-deficient soft tissue sarcoma NOS	Sarcoma
teratoma	Teratoma
infantile fibrosarcoma	Sarcoma
acinar cell carcinoma	Pancreatic Cancer
pineal parenchymal tumor	Brain Cancer
rosette forming glioneuronal tumor	Brain Cancer
juvenile myelomonocytic leukemia	JMML
myeloid neoplasm NOS	AML
alveolar soft part sarcoma	Sarcoma
follicular neoplasm	Thyroid Cancer
ganglioglioma	Brain Cancer
undifferentiated hepatic sarcoma	Liver Cancer
atypical teratoid/rhabdoid tumor	Rhabdoid
acute undifferentiated leukemia	Leukemia
retinoblastoma	Retinoblastoma
germ cell tumor	Germ Cell Tumor
inflammatory myofibroblastic tumor	inflammatory myofibroblastic tumor
meningioma	Brain Cancer
chronic myelogenous leukemia (S02), acute lymphoblastic leukemia (S01)	CML
neoplasm (uncertain whether benign or malignant)	Unknown
Sertoli-Leydig cell tumor, retiform	Germ Cell Tumor
epithelioid sarcoma	Sarcoma
fibromatosis	Fibromatosis
rhabdoid tumor	Rhabdoid
myeloproliferative neoplasm	Unknown
adrenocortical cancer	Adrenal Cancer
lipoblastomatosis	Lipoblastomatosis
adrenocortical adenoma	Adrenal Cancer
pleomorphic myxoid liposarcoma	Liposarcoma

sclerosing epithelioid fibrosarcoma	Sarcoma
embryonal tumor with multilayered rosettes	Sarcoma
endometrial stromal sarcoma	Sarcoma
epithelioid hemangioendothelioma	Hemangioendothelioma
nasopharyngeal carcinoma	Head and Neck Cancer
craniopharyngioma	Brain Cancer
NUT midline carcinoma	Head and Neck Cancer
angiosarcoma	Sarcoma
gliomatosis cerebri	Brain Cancer
neurofibromatosis type 1	Neurofibroma
neurofibroma	Neurofibroma
leukemia	Leukemia
thymic carcinoma	Thyroid Cancer
clear cell carcinoma of cervix	Cervical Cancer
PEComa	PEComa
undifferentiated spindle cell sarcoma	Sarcoma
myofibromatosis	Fibromatosis
pleuropulmonary blastoma	Lung Cancer
melanotic neuroectodermal tumor	melanotic neuroectodermal tumor
neuroendocrine carcinoma	neuroendocrine carcinoma

Appendix C

GSEA of enriched vs induced genes

See next page.

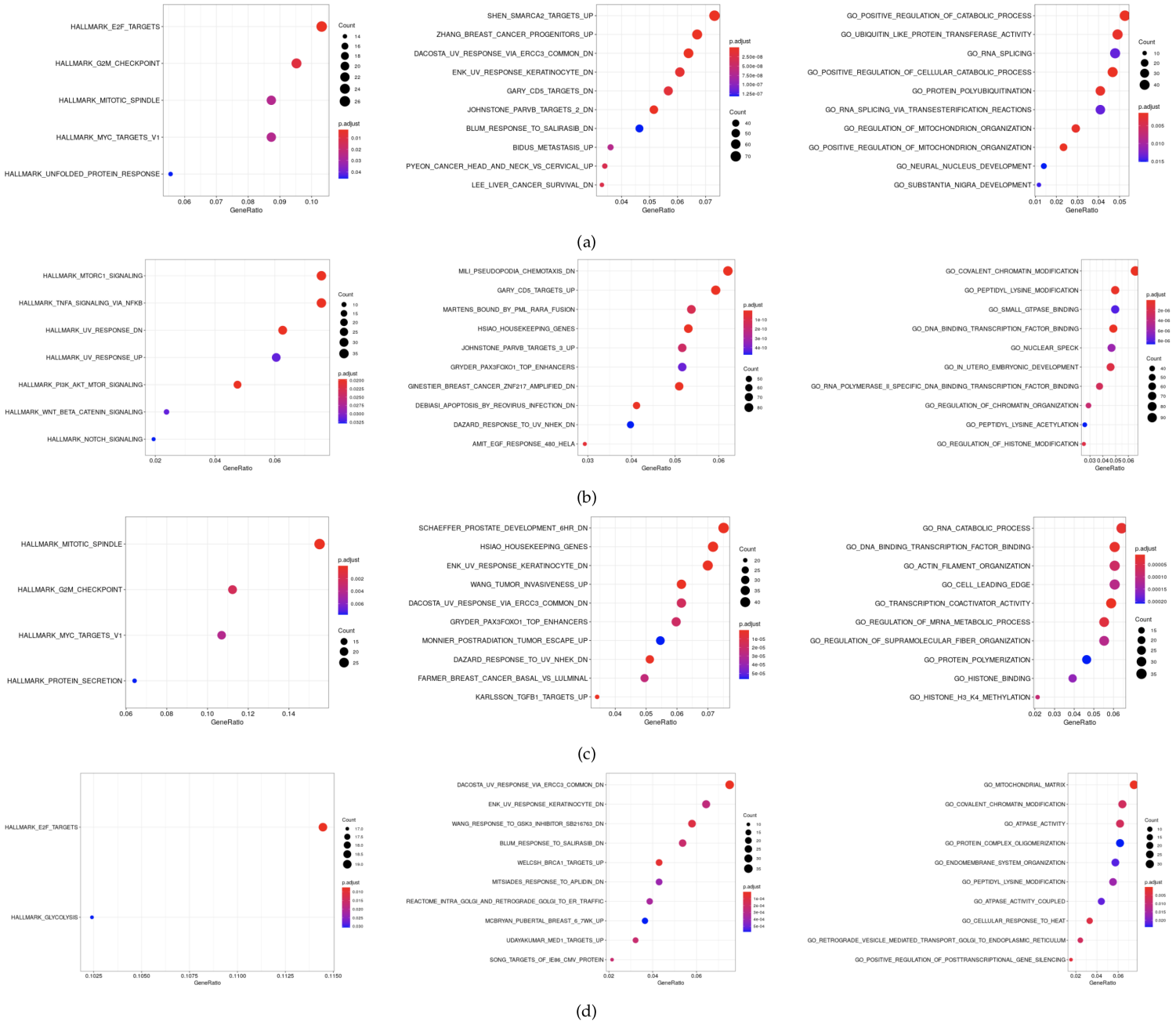


Figure C.1: Gene set enrichment analysis in Hallmark, Curated and GO gene sets of the induced and repressed genes. (a) and (b) results of the whole cell RNAseq data, (c) and (d) results of the 24 hours proteomics data.