Functional compensators of epigenetic modifiers as targetable cancer vulnerabilities

Dressler, Lisa

Awarding institution:
King's College London

Functional compensators of epigenetic modifiers as targetable cancer vulnerabilities

Lisa Dressler

King's College London

and

The Francis Crick Institute

PhD Supervisors: Francesca Ciccarelli & Rebecca Oakey

A thesis submitted for the degree of

Doctor of Philosophy

King's College London

December 2021

Declaration

I Lisa Dressler confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Due to an increasing number of cancer sequencing screens, our knowledge of genes whose somatic alterations drive cancer initiation and progression (cancer drivers) has vastly expanded over the past 15 years. Sequencing screens have enabled a comprehensive collection of cancer drivers across tissues. This has revealed their characteristic evolutionary properties; lower gene duplicability, early evolutionary origin, ubiquitous RNA and protein expression, numerous miRNA interactions, participation in complexes by encoded proteins and a central, connected, and inter-connected position in the protein-protein interaction network. Once cancer drivers are identified, understanding the effect of their alterations forms the foundation of personalized cancer therapies. Targeting cancer-specific vulnerabilities, such as inhibiting functional compensators of genes whose function is lost in cancer cells, minimises the side effects in non-cancer cells.

Considering the continuous development of scientific knowledge in this area, the first part of this thesis addresses the improvement and expansion of the collection and characterisation of cancer drivers included in the Network of Cancer Genes resource. It comprises an up-to-date collection of 3,347 altered genes driving cancer and 95 genes driving clonal expansion in non-malignant tissues. In addition to confirming their evolutionary properties, we find that cancer drivers are more essential and less robust against damaging germline alterations. We reveal distinctive properties of different driver gene categories; known and predicted cancer drivers, drivers with coding and noncoding alterations, and genes driving cancer and non-cancer clonal expansion.

To place cancer driver properties in a wider context, the second part of this thesis characterises evolutionary properties of functional gene groups in health and disease. We integrate nine evolutionary properties into a single score using a random forest classifier. This divides 25 biological pathways encompassing 10,334 genes into three groups. Genes in high-scoring pathways perform basic cell functions and are enriched in tumour suppressors and core essential genes. In contrast, genes in low-scoring pathways contribute to organ-specific functions and are enriched in recessive Mendelian disease genes. Intermediate-scoring pathways contribute to metabolism, development, and immune system. The

integrated analysis of gene evolutionary properties using a principal component analysis prioritises a subgroup of predicted cancer genes for further validation.

We use a subset of evolutionary gene properties to predict functional compensation between gene pairs. To this aim, we develop a computational prediction method that combines genetic sequence conservation, engagement in the same protein complex and context dependent gene essentiality in cancer cell lines. We show that epigenetic modifiers are enriched in paralog pairs, genes encoding proteins that engage in complexes and context dependent essential genes and are frequently lost in cancer. Consistent with this, they are enriched in predicted functional compensator pairs, making them interesting therapeutic targets. Thus, we focus validation on epigenetic modifiers. Using CRISPR Cas9 mediated gene knockout, we validate synthetic lethality between *GATA2* and *GATA3*, as well as context dependent synthetic lethality between *TBL1X* and *TBL1XR1*.

In summary, this thesis explores cancer gene properties, analyses them in the context of broader functional groups and prioritises new cancer genes for validation. It identifies an enrichment of epigenetic modifiers in potential paralog synthetic lethal interactions and validates synthetic lethality between two gene pairs.

Acknowledgements

I would like to thank my supervisors Prof. Francesca Ciccarelli and Prof. Rebecca Oakey and the members of my thesis committee Dr. Paola Scaffidi, Prof. Victoria Sanz-Moreno, and Dr. Frank Uhlmann. I would also like to thank all members, past and present, of the Cancer Systems Biology Laboratory and the Epigenetics Laboratory. I am thankful for the support of all funding bodies involved in supporting the work presented in this thesis. Finally, I would like to thank my family and friends for their continuous support.

Table of Contents

Table of figures

List of tables

Abbreviations

Abbreviation	Definition
ATP	Adenosine triphosphate
ATPase	Adenosine triphosphatase
BLAT	Basic-local-alignment-search-tool-like alignment tool
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DAISY	DAta mIning SYnthetic lethality
DNA	Deoxyribonucleic acid
DMEM	Dulbeco's Modified Eagle's Medium
DTT	Dithiothreitol
EP score	Evolutionary property score
FCS	Foetal calf serum
FDA	Food and Drug Administration
FDR	False discovery rate
FPKM	Fragment per kilobase per million
gRNA	Guide RNA
H3K18ac	Histone 3 lysine 18 acetylation
H3K27me3	Histone 3 lysine 27 trimethylation
H3K27ac	Histone 3 lysine 27 acetylation
H3K29	Histone 3 lysine 29
H3K36me3	Histone 3 lysine 36 trimethylation
H3K4	Histone 3 lysine 4
H3K79	Histone 3 lysine 79
H3K79ac	Histone 3 lysine 79 acetylation
H3K79me2	Histone 3 lysine 79 dimethylation
H3K79me3	Histone 3 lysine 79 trimethylation
H3K9ac	Histone 3 lysine 18 acetylation
H3K9me3	Histone 3 lysine 9 trimethylation
H4K16ac	Histone 4 lysine 16 acetylation
ICGC	International Cancer Genome Consortium
ISLE	Identification of clinically relevant Synthetic LEthality

ISWI	Imitation SWItch
LOEUF	Loss-of-function observed/expected upper bound fraction
miRNA	Micro RNA
MiSL	Mining Synthetic Lethality
mRNA	Messenger RNA
NADPH	Nicotinamide adenine dinucleotide phosphate
nTC	Non target control
NCG	Network of Cancer Genes
PARP	Poly adenosine diphosphate ribose polymerase
PARylate	Poly adenosine diphosphate ribosylate
PBS	Phosphate buffered saline
PCA	Principal component analysis
PCR	Polymerase chain reaction
PPIN	Protein-protein interaction network
RNA	Ribonucleic acid
RNAi	RNA interference
RNase	ribonuclease
RPKM	Reads per kilobase million
RPMI	Roswell Park memorial institute
shRNA	Short hairpin RNA
siRNA	Small interfering RNA
SLant	Synthetic Lethal analysis via network topology
SNV	Single nucleotide variant
SWI/SNF	SWItch/Sucrose Non-Fermentable
TAE	Tris acetate ethylenediaminetetraacetic acid
TBS	Tris buffered saline
TBST	Tris buffered saline tween
TCGA	The Cancer Genome Atlas
TPM	Transcripts per million
UTR	Untranslated region
WT	Wild type

Chapter 1. Introduction

1.1 Cancer, a result of malignant clonal expansion

1.1.1 Somatic alterations leading to cancer

Despite great advances in diagnosis and therapy, cancer is one of the leading causes of death worldwide. In 2019, 9.3 million deaths were caused by cancer worldwide, with five million deaths recorded for patients under the age of 70 (World Health Organization, 2020). This makes cancer the first or second leading cause of death before the age of 70 in 112 out of 183 countries (Sung et al., 2021). Over the next 20 years, cancer mortality is expected to rise further, increasing to over 16 million cases in 2040 (Ferlay J, 2020). This increase will most drastically affect countries with a low human development index and a high increase in population, leading to a doubling of cancer incidence (Ferlay J, 2020).

To address this issue, a detailed understanding of cancer origin and progression is necessary. Cancer is caused by somatic alterations which constantly accumulate during the course of our lives. They are caused by endogenous damage, such as faulty deoxyribonucleic acid (DNA) replication or spontaneous reactions of the DNA with water or reactive oxygen species, as well as exogenous factors such as ionizing radiation, ultraviolet radiation, chemicals and environmental stress (Chatterjee and Walker, 2017). Some of these alterations affect genes essential for cell survival and are therefore under negative selection pressure (Bartha et al., 2018; Stratton et al., 2009). The majority of alterations accumulate in the cell without any effect, either because they are silent mutations or because they do not affect protein functions (Martincorena, 2019; Martincorena and Campbell, 2015; Stratton *et al.*, 2009). Finally, those alterations that confer a growth advantage on the cell lead to clonal cell expansion (Martincorena, 2019). For the transformation of normal clones into cancer, a series of essential criteria need to be met as defined by Hanahan and Weinberg (Hanahan and Weinberg, 2000). Cells need to resist cell death and achieve replicative immortality. In addition, they need to evade growth suppression and sustain proliferative signalling. Finally, growth needs to be sustained by induced

angiogenesis, and further spread enabled through invasion and metastasis. This concept has been expanded by two additional emerging hallmarks, the avoidance of immune destruction and the deregulation of cellular metabolism to support rapid cell growth (Hanahan and Weinberg, 2011). Furthermore, two enabling characteristics were also identified, namely genomic instability and tumour promoting inflammation. The requirements for the development of cancer are met by the accumulation of genetic alterations (Gerstung et al., 2020; Shendure and Akey, 2015; Stratton *et al.*, 2009). DNA alterations contributing to cancer initiation and progression are called cancer driver alterations and genes that are altered by these alterations are called cancer driver genes (Figure 1-1A).

All DNA alterations are subject to selective pressure (Cairns, 1975; Greaves and Maley, 2012). In contrast to species evolution, where germline mutations are predominantly impacted by negative selection, tumours mainly evolve due to positive selective pressure on somatic alterations (Martincorena et al., 2017; Ostrow et al., 2014). Positive selection and the resulting recurrence of driver alterations is therefore used to distinguish cancer drivers from non-driver alterations called passenger alterations (Martinez-Jimenez et al., 2020). Passengers are alterations which themselves have a small or no impact on cell growth. They are detected in cancer because they hitchhike on the clonal expansion caused by driver alterations. In contrast, positive selective pressure on driver alterations leads to accumulation of mutations in a driver gene, clusters of alterations in certain parts of the gene, bias towards alterations with high functional impact and alterations in a certain trinucleotide context (Martinez-Jimenez *et al.*, 2020). This recurrence of cancer drivers is the most common criterion used to identify driver genes (Repana et al., 2019). Computational tools such as MutSig (Lawrence et al., 2013), MuSiC (Dees et al., 2012) and OncodriveCLUST (Tamborero et al., 2013) or combinations of these and further tools (Bailey et al., 2018) have fine-tuned this concept by estimating positive selection of alterations above background mutation rates. All of these computational methods rely on a large resource of cancer sequencing data to have sufficient statistical power. To this aim, thousands of cancer patient genomes have been sequenced by large sequencing initiatives such as The

Cancer Genome Atlas (TCGA) (TCGA Research Network, 2021), the International Cancer Genome Consortium (ICGC) (Zhang et al., 2019), the 1+ Million Genomes Initiative (Beyond 1 Million Genomes, 2021) and the 100,000 genomes project (Caulfield et al., 2017). For example, the TCGA project has sequenced and molecularly characterized over 20,000 cancers and matched normal samples from 33 different cancer types. Since the power to detect positively selected alterations correlates with the cohort size (Bailey *et al.*, 2018; Dressler et al., 2021; Repana *et al.*, 2019), the addition of further sequenced genomes will help to identify additional cancer drivers mutated at lower frequency and in rare cancer types.

The number of genes under positive selection in a cancer depends on the cancer type (Martincorena *et al.*, 2017). With an average of four coding alterations under positive selection identified in tumours, numbers range from one alteration in thyroid and testicular cancers to as high as ten alterations in endometrial and colorectal cancers. Interestingly, only approximately half of these alterations under positive selection occur in known cancer driver genes, confirming that there is still a need for further investigation into novel cancer drivers (Martincorena *et al.*, 2017).

Cancer drivers are classified into two groups based on their mechanism of action. Genes driving cancer upon gain-of-function alterations, meaning alterations that enhance the biological function of their encoded protein, are called oncogenes (Klein and Klein, 1985). They are usually involved in pathways promoting cell growth, and their gain of function has a dominant phenotype. Their alterations often occur through activating mutations, amplification events, chromosomal translocations (Lodish et al., 2000) or aberrant epigenetic modification leading to overexpression (Egger et al., 2004). Conversely, tumour suppressors are affected by loss-of-function alterations. These include deletions, damaging mutations and combinations of both and result in a reduced or abolished protein function (Caldas and Venkitaraman, 2001). In addition, loss of function can occur through epigenetic downregulation, non-coding ribonucleic acids (RNAs) or defects in transcription processes (Kazanets et al., 2016). Typical functions of tumour suppressors include suppression of cell division, induction of apoptosis,

DNA damage repair and inhibition of metastasis (Sun and Yang, 2010). The role of some drivers is still unknown, and other drivers perform differing roles depending on the context (Shen et al., 2018). For example, the well-known tumour suppressor *TP53* can also act as an oncogene (Soussi and Wiman, 2015). Already by 1993 it was shown that the expression of mutant *TP53* leads to enhanced cancer cell growth both in cell lines and mouse models lacking endogenous *TP53* expression (Dittmer et al., 1993). In addition, alterations in the Notch signalling pathway influence cell survival and cell fate. Depending on the cancer type, frequent loss-of-function as well as gain-of-function alterations have been observed in *Notch1*, *Notch2* and *Notch3*, indicating a dual role as tumour suppressor or oncogene (Aster et al., 2017).

Traditionally, the search for cancer genes has focused on alterations that affect the protein coding part of genes. With whole genome sequencing becoming more accessible, the non-coding regions of the human genome are investigated more closely (Elliott and Larsson, 2021; Liu et al., 2021; Weinhold et al., 2014). Non-coding driver alterations include mutations in enhancers, promoters, 5' or 3' untranslated regions (UTRs), splice sites, regulatory RNA elements or chromatin conformation regulatory regions (Liu *et al.*, 2021). These elements present technical challenges, such as significantly lower sequencing coverage of CG-rich elements at promoters or 5'UTR regions (Wang et al., 2011). In addition, experimental validation of non-coding drivers through differential gene expression or other experimental support is still lacking (Liu *et al.*, 2021). Despite these challenges, non-coding drivers further contribute to our understanding of cancer and could point towards new therapeutic strategies.

1.1.2 Personalized therapy as a response to tumour heterogeneity

Through substantial advances in cancer sequencing studies, it has become clear that considering cancer as "one disease" is misleading. Instead, each cancer sample represents a unique case, both regarding its mutational profile and the timing at which alterations are acquired (Bailey *et al.*, 2018; Cusnir and Cavalcante, 2012; Gerstung *et al.*, 2020; Nowell, 1976). This inter-tumour

heterogeneity results in variable treatment outcomes depending on underlying alterations. Therefore, personalized cancer treatment is a promising avenue to improve therapy outcomes by specifically addressing each tumour's weak points (Ameratunga et al., 2020; Bailey et al., 2014; Dohner et al., 2021; Dumbrava and Meric-Bernstam, 2018; Krzyszczyk et al., 2018; Nakamura et al., 2021; Shin et al., 2017; Xie et al., 2020; Yuan et al., 2019). The first success story of personalized therapy is the successful stratification of breast cancer patients based on their *HER2* status. Only patients with *HER2* amplification or overexpression respond to trastuzumab, a monoclonal antibody targeting *HER2* (Vogel et al., 2002). Another example involves tyrosine kinase inhibitors which are an effective treatment for chronic myelogenous leukaemia patients harbouring a translocation of chromosome 9 and 22. This alteration results in the fusion protein Bcr-Abl, which drives cancer formation through upregulation of tyrosine kinase activity (O'Brien et al., 2003). More recently, poly adenosine diphosphate ribose polymerase (PARP) inhibitors such as olaparib were shown to be especially effective against ovarian, breast and prostate cancers with *BRCA1* or *BRCA2* mutations (Fong et al., 2009; Fong et al., 2010). As the field of personalized treatment is expanding rapidly, these are only few exemplary success stories of targeted cancer treatment.

Besides inter-tumour heterogeneity, a further challenge for successful cancer therapy is intra-tumour heterogeneity, describing a tumour's heterogeneous composition of different clones (Ramon et al., 2020). In primary tumours, intra-tumour heterogeneity is highly variable within and between cancer types (McGranahan and Swanton, 2017). It is typically low in tissues exposed to exogenous damaging factors, for example lung adenocarcinoma and melanoma, and increases upon cancer therapy. Intra-tumour heterogeneity also plays an important role for metastasis formation, since the additional acquisition of new alterations initiates the spread to distant sites (Gerlinger et al., 2012)(Figure 1-1A). In a multi-region sampling study of primary and metastatic renal carcinomas, only 34% of all mutations were present in all regions of the primary tumour and all metastases (Gerlinger *et al.*, 2012). In concordance, a study of 1,621 tumours found that less than 5% of tumours were composed of only one homogeneous

clone (Dentro et al., 2021). Therefore, it is crucial to target personalized treatment to clonal alterations in order to avoid competitive release of resistant subclones (McGranahan and Swanton, 2017).

1.1.3 Clonal expansion in non-cancer tissues

Clonal expansion is not always malignant and has been observed in numerous non-cancer tissues (Kakiuchi and Ogawa, 2021; Martincorena, 2019; Wijewardhane et al., 2020) (Figure 1-1A). For example, the normal adult human epidermis is a patchwork of mutant clones which constantly compete for space to expand and survive (Martincorena et al., 2015). Thus, analogous to cancer driver genes, somatic mutations of healthy driver genes drive non-malignant clone formation based on predominantly positive selection pressure (Martincorena *et al.*, 2017). Non-malignant mutant clones may even have an anti-tumorigenic effect by outcompeting early malignant neoplasms (Colom et al., 2021). Further examples of tissues where non-malignant clonal expansion occurs include the oesophagus (Martincorena et al., 2018; Yokoyama et al., 2019), liver (Brunner et al., 2019), bladder (Lawson et al., 2020), colon (Lee-Six et al., 2019), and endometrium (Lac et al., 2019; Moore et al., 2020; Suda et al., 2018) (Figure 1-1B). Of note, some studies also investigated chronically inflamed tissues such as endometriotic endometrium (Anglesio et al., 2017; Lac et al., 2018; Suda *et al.*, 2018), colon affected by inflammatory bowel disease (Olafsson et al., 2020) or cirrhotic liver (Brunner *et al.*, 2019; Zhu et al., 2019) (Figure 1-1B). These tissues usually have a higher mutational burden than healthy tissues and represent pre-malignant states that are predisposed to future cancer development. Single alterations of cancer driver genes may not be enough to initiate cancer. For example, up to 19% of adult bladder epithelial cells harbour cancer driver alterations (Lawson *et al.*, 2020). In addition, the cancer driver genes *NOTCH1* and *PPMID* are mutated more frequently in heathy oesophagus than oesophageal carcinoma (Yokoyama *et al.*, 2019). As described in the previous section, the accumulation of several cancer drivers is often needed to initiate malignant tumour growth.

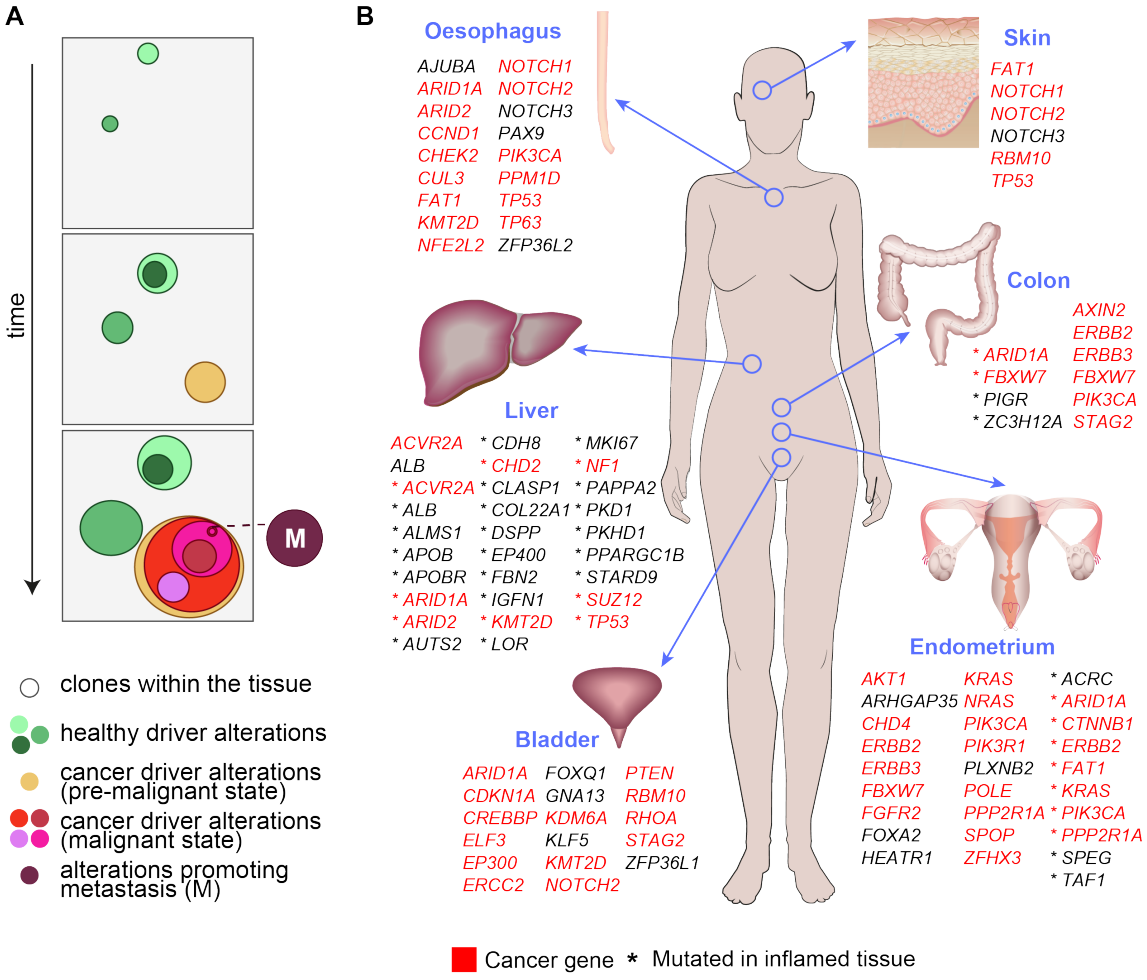


Figure 1-1 Cancer and healthy drivers.

A) Non-malignant and malignant clonal expansion over time. Over time, cells acquire growth promoting alterations which lead to clonal expansion. These can be healthy driver alterations or cancer driver alterations. While few cancer driver alterations may lead to a pre-malignant state, the accumulation of cancer drivers leads to tumour formation. Tumours may be heterogeneous and metastasize. **B)** Mutated genes under positive selection or frequently mutated in the human adult tissue. Healthy drivers were collected from 13 studies. Some of the drivers, indicated by (*), were identified in pre-malignant inflamed tissue. Many healthy drivers are also cancer drivers. Figure adapted from Wijewardhane et al. (Wijewardhane *et al.*, 2020).

1.2 Evolutionary properties of genes

1.2.1 Evolutionary properties of cancer genes

With the growing knowledge of cancer genetics, it becomes increasingly difficult to keep an overview of cancer drivers. This issue has been addressed by several cancer driver databases with diverse areas of interest. For example, some databases focus on driver alterations (Ainscough et al., 2016; Cerami et al., 2012; Chakravarty et al., 2017; Liu et al., 2020; Tamborero et al., 2018; Tate et al., 2019), while others collect different categories of driver genes. These may be driver genes with strong indication of involvement in cancer (Futreal et al., 2004), with a specific function (Liu et al., 2017; Zhao et al., 2016), in a specific cancer type (Agarwal et al., 2016) or drivers collected through a non-curated literature-mining approach (Lever et al., 2019). The IntOGen framework predicts cohort-specific cancer driver genes and their mechanism of action (Martinez-Jimenez *et al.*, 2020). It is based on a method called boostDM, which analyses point mutations from cancer sequencing data using seven driver prediction methods.
None of these databases provide a well curated, comprehensive overview of all currently known cancer genes. To fill this gap, the Ciccarelli lab is maintaining a database of cancer genes called the Network of Cancer Genes (NCG, http://network-cancer-genes.org). NCG provides a comprehensive list of well-characterised canonical cancer genes as well as candidate cancer genes obtained through a literature review (An et al., 2016; An et al., 2014; D'Antonio et al., 2012; Syed et al., 2010). In addition, NCG annotates gene evolutionary properties, including gene duplicability, evolutionary origin, microRNA (miRNA) interactions, RNA and protein expression as well as protein interactions and participation in protein complexes by the encoded protein. These properties are shaped by a gene's evolutionary path and its biological role. Cancer genes are characterized by lower duplicability in the human genome (Rambaldi et al., 2008), early evolutionary origin (D'Antonio and Ciccarelli, 2011; Domazet-Loso and Tautz, 2008; 2010), are more frequently targeted by miRNAs (An *et al.*, 2016) and are ubiquitously expressed (An *et al.*, 2016). They encode proteins that engage in complexes and are highly connected (hubs), inter-connected and

central in the human protein-protein interaction network (PPIN) (An *et al.*, 2016). Within the group of cancer genes, tumour suppressors originated earlier in evolution than oncogenes (D'Antonio and Ciccarelli, 2011; Domazet-Loso and Tautz, 2010). While tumour suppressors tend to maintain a single copy status, oncogenes more often have duplicates (An *et al.*, 2016). The distinct properties of cancer genes have proved useful in detecting patient specific cancer drivers through machine learning (Mourikis et al., 2019; Nulsen et al., 2021).

1.2.2 Evolutionary properties of gene groups in health and disease

While cancer is caused by somatic alterations, Mendelian diseases are caused by alterations in the germline. Mendelian disease genes also have characteristic properties: they originated early in evolution (Domazet-Loso and Tautz, 2008), but the position of their encoded proteins in the PPIN is variable. For example, Mendelian disease genes with embryonic lethal orthologs in mouse encode central protein hubs, while disease genes with viable knockout orthologs encode proteins at the periphery of the network (Dickerson et al., 2011; Goh et al., 2007). In addition, genetic disorders caused by mutations in enzyme-encoding genes are often recessive, while germline alterations of transcription factors are associated with dominant inheritance (Jimenez-Sanchez et al., 2001).

Independent of disease association, biological processes often have similar properties. For example, fundamental biological processes such as messenger RNA (mRNA) processing or protein transport have changed very little since early single cell lifeforms, and are mainly performed by highly connected proteins (hubs) originating early in evolution (Szedlak et al., 2016). Furthermore, proteins involved in DNA repair are more connected, located at the PPIN centre and tend to interact directly with each other (Li and Zhang, 2017). Proteins involved in the regulation of the circadian rhythm (Castellana et al., 2018) or the neural crest (Sauka-Spengler et al., 2007) also originated early in evolution. In contrast, proteins that originated more recently in evolution are involved in species- or cell type-specific functions such as immune system or olfactory related pathways, and are at the periphery of the PPIN (Szedlak *et al.*, 2016). This allows for a

higher evolutionary rate, meaning the ratio of nonsynonymous to synonymous mutations, to fine-tune their function and optimize fitness at cellular or organism level.

Genes that are essential for survival of human cells have characteristic properties that resemble cancer gene properties. They predominantly originated early in evolution (Chen et al., 2020; Chen et al., 2012), are not duplicated in the human genome and are broadly expressed across tissues (Chen *et al.*, 2020; Wang et al., 2015). They also encode proteins which are highly connected in the PPIN (Chen *et al.*, 2020; Wang *et al.*, 2015) and participate in protein complexes (Hart et al., 2014). In addition, essential genes and cancer drivers are enriched in similar biological processes. Both gene classes accomplish basic functions, such as gene transcription, DNA replication and repair and cell cycle regulation (Chen *et al.*, 2020; Hart *et al.*, 2014; Repana *et al.*, 2019).

The similarities of these two gene classes are reflected in the topology of the PPIN. Initial analysis of the yeast PPIN revealed its scale-free topology (Jeong et al., 2001), which is characterised by few highly connected hubs and a majority of less connected nodes at the periphery of the network. The main advantage of the scale-free topology is a high degree of tolerance towards random errors in the network. This tolerance is based on the network flexibility, achieved through the highly connected hubs that maintain connectivity even when errors occur in the periphery of the network (Albert et al., 2000). However, disturbance of the hubs leads to failure of the whole system. Further investigations and comparison of protein interactions in yeast and human concluded that the PPIN is better described by the highly optimized tolerance network topology (Hase et al., 2009). In addition to highly vulnerable, essential hubs and error-tolerant peripheral nodes, it includes highly inter-connected middle-degree nodes forming the backbone of the network. These are enriched in genes causing inherited diseases (Feldman et al., 2008) and cancer drug targets (Hase *et al.*, 2009).

1.2.3 Interdependence of evolutionary properties

To ensure the compatibility of a mutation in the protein coding part of the genome with the interactors of the encoded proteins, the interactors need to coevolve. This becomes less likely with increasing numbers of interactors (Fraser et al., 2002). Therefore, the most central and connected proteins of the yeast PPIN, which also originated earliest in evolution, have the slowest evolutionary rates (Fraser *et al.*, 2002). This was also confirmed for the human gene regulatory network, a network constructed by integration of mRNA co-expression, protein-protein interactions, protein complexes and comparative genomics datasets (Szedlak *et al.*, 2016). In contrast, peripheral nodes appeared later in evolution, but evolve more rapidly (Szedlak *et al.*, 2016). Thus, the human gene regulatory network consists of slowly evolving, old, central hubs and peripheral, young, rapidly evolving genes.

Besides evolutionary age, the number of targeting miRNAs positively correlates with the protein connectivity in the human PPIN (Liang and Li, 2007b). Within the group of highly targeted hubs, miRNA targeting negatively correlates with the PPIN clustering coefficient. This indicates a greater need for regulation of hubs connecting functional modules (intermodular hubs) than hubs within one functional module (Liang and Li, 2007b). Genes expressed across many tissues are also enriched in miRNA targets (Yu et al., 2007). Gene duplicability has changed throughout evolution. In yeast, worm and fly, genes encoding protein hubs are usually preserved as a single copy (Hughes and Friedman, 2005; Makino et al., 2009; Papp et al., 2003; Prachumwat and Li, 2006; Yang et al., 2003). A possible explanation may be the susceptibility of evolutionary old protein hubs to dosage modification, which disrupts the stoichiometry of protein-protein interactions (Veitia, 2002; 2004). This limitation was overcome later in evolution through whole genome duplication in vertebrates, tissue-specific expression (Freilich et al., 2005) or regulation through miRNAs (D'Antonio and Ciccarelli, 2011). Therefore, protein hubs originating in vertebrates or later can be encoded by both single copy or duplicated genes (Liang and Li, 2007a; Liao and Zhang, 2007; Makino *et al.*, 2009; Rambaldi *et al.*, 2008).

Overall, it becomes apparent that evolutionary properties are highly interconnected. In addition, gene groups in health and disease are characterized by distinctive evolutionary properties (Figure 1-2).

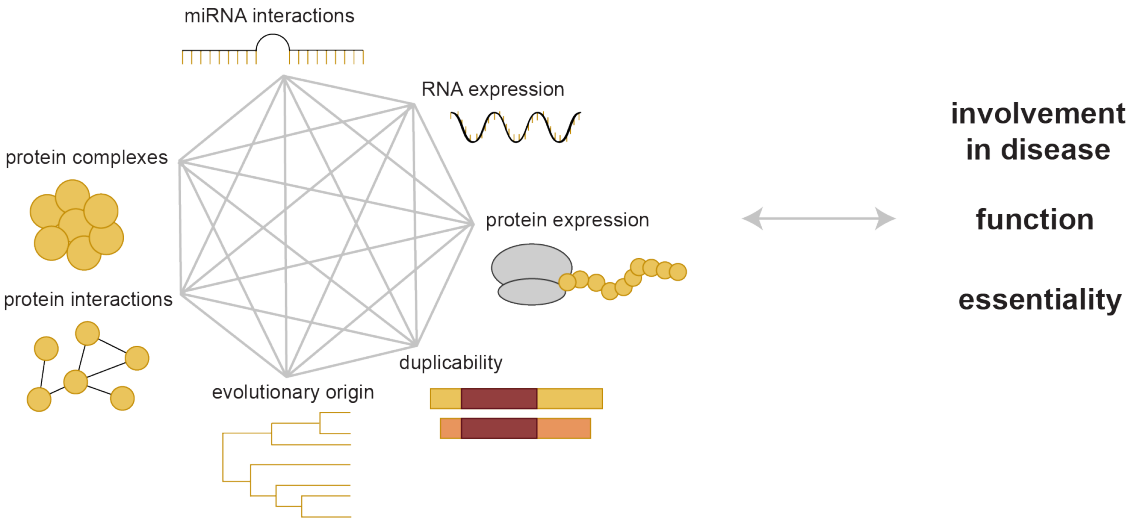


Figure 1-2 Evolutionary properties and involvement in disease, function, and essentiality.

Gene evolutionary properties are interconnected, and are also connected to gene involvement in disease, gene function and gene essentiality.

1.3 Functional compensation between genes

1.3.1 Functional compensation and synthetic lethality

The topology of the human PPIN makes it robust to perturbations in the form of mutations. A further increase in robustness is achieved through gene duplication (Kafri et al., 2006). Genetic duplicates also work together in fine-tuning expression levels and filtering noise from transcriptional pathways, thus leading to more sophisticated regulatory possibilities (Kafri *et al.*, 2006). In addition, sub-functionalization, meaning mutations inducing slight differences in expression patterns or function, can lead to an evolutionary advantage and retention of duplicates (Force et al., 1999).

Robustness towards deleterious alterations through gene duplication is based on their functional compensation. Two genes with overlapping functions may be able

to compensate for each other's loss, enabling the cell to survive without one of them (Rehman et al., 2010). However, if their function is essential, losing both will lead to cell death. The genetic interaction whereby the loss of either gene of a duplicate pair is viable but the loss of both leads to cell death is a prominent example of synthetic lethality. More generally, a synthetic lethal interaction between two or more genes describes the phenomenon whereby the perturbation of either gene alone is viable, but a combination of perturbations is not (O'Neil et al., 2017).

Many examples of synthetic lethality based on gene duplicates have been discovered. For example, the interchangeable subunits of the switch/sucrose non-fermentable (SWI/SNF) complex (Mashtalir et al., 2018), *ARID1A*/*ARID1B* (Helming et al., 2014) and *SMARCA2*/*SMARCA4* (D'Antonio et al., 2013; Hoffman et al., 2014; Oike et al., 2013), are synthetic lethal partners. Further synthetic lethal pairs include the cohesin subunits *STAG1*/*STAG2* (Benedetti et al., 2017; van der Lelij et al., 2017), histone methyltransferases *EZH1*/*EZH2* (Honma et al., 2017) and *KMT2A*/*KMT2B* (Ernst et al., 2016) and histone acetyltransferases *CREBBP*/*EP300* (Ogiwara et al., 2016). *ME2* targeting selectively kills cancer cells with *ME3* mutations, as one functional gene is essential for nicotinamide adenine dinucleotide phosphate (NADPH) regeneration (Dey et al., 2017). *MAGOH* and *MAGOHB*, core members of the splicing-dependent exon junction complex (Viswanathan et al., 2018) and the histone chaperones *ASF1A*/*ASF1B* (De Kegel et al., 2021) are further examples of duplicate synthetic lethality. *COPS7A*/*COPS7B* which are involved in the ubiquitin conjugation pathway as part of the signalosome complex are also synthetic lethal duplicates (De Kegel *et al.*, 2021). The phosphatases *DUSP4*/*DUSP6* are synthetic lethal interactors in cell lines with overactivation of the MAPK pathway (Ito et al., 2021).

Synthetic lethality through loss of function is not limited to two duplicates. For example, a yeast triple-mutant screen identified synthetic lethal interactions between gene triplicates involved in chromosome regulation, such as the histone chaperones *ASF1* and *CAC1* and the SWI/SNF complex component *RDH54*

(Haber et al., 2013). In cancer cell lines, loss of *DNMT3A*, *DNMT3B* and *DNMT1* alone is viable, but loss of all three leads to cell death (D'Antonio *et al.*, 2013).

In addition to duplicate synthetic lethality, single alterations in parallel pathways may be viable, but lethal in combination. For instance, the previously mentioned targeted therapy with PARP inhibitors, which is especially effective in *BRCA1* and *BRCA2* mutated cancers, is based on the synthetic lethality between alterations in two parallel DNA damage repair pathways (Fong *et al.*, 2009; Fong *et al.*, 2010). *PARP1* and *PARP2* contribute to the base excision repair mechanism for single-strand DNA breaks. They bind and poly adenosine diphosphate ribosylate (PARylate) damaged DNA, leading to recruitment of DNA repair effectors (Lord and Ashworth, 2017). Alternatively, single-strand breaks can be converted to double-strand breaks, which enables repair through homologous recombination mediated through *BRCA1* and *BRCA2*. The loss of either pathway for DNA damage repair is not lethal, however, loss of both leads to cell death through accumulation of DNA damage, especially after chemotherapy treatment. Other DNA damage repair pathways may represent a rich source of further synthetic lethal interactions (Brown et al., 2017; Das et al., 2019).

Synthetic dosage lethality describes the lethal effect of simultaneous under-expression of one gene and over-expression of another gene (Kroll et al., 1996). For example, in yeast, over-expression of *CTF13* is incompatible with kinetochore mutants, and over-expression of *ORC6* causes synthetic dosage lethality in combination with mutations in the replication pathway. A further example is the conserved genetic interaction between *PLK1* loss and over-expression of *CKS1B* in yeast and human breast cancer cells (Reid et al., 2016). This concept was further extended to include low, medium and high expression levels of two genes, whereby any combination of expression patterns of two genes may result in a growth advantage or disadvantage to the cell (Magen et al., 2019). While synthetic lethality is the most extreme form of synthetic interactions, genetic combinations which result in non-lethal, but impaired growth are called synthetic sickness (Nijman, 2011).

1.3.2 Experimental identification of synthetic lethal interactions

The concept of synthetic lethality was first described in *drosophila melanogaster* in 1922, where only the combination of certain genetic variants was found to be lethal (Bridges, 1922). Its name was coined more than 20 years later (Dobzhansky, 1946), with the Greek origin of the word 'synthetic' meaning 'combination of two entities to form something new' (Nijman, 2011).

High throughput investigation of synthetic lethal interactors was first performed in *saccharomyces cerevisiae*, where genetic interactions between two genes were investigated through systematic query of double mutants (Tong et al., 2001; Tong et al., 2004). Comparisons between *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* revealed a conservation of merely 29% of genetic interactions, indicating that evolutionary distance and genetic context have a significant impact on interactions. Nevertheless, strong genetic interactions in yeast are more likely to be conserved in humans (Srivas et al., 2016). Therefore, yeast screens are helpful to guide synthetic lethality investigations in humans. An increased throughput of over 23 million double mutants in *saccharomyces cerevisiae* identified comprehensive genetic interaction maps (Costanzo et al., 2016). This network revealed a clustering of genetic interactions within similar cellular functions, allowing the prediction of gene function based on interaction clusters.

High throughput RNA interference (RNAi) screens using immortalised or cancer cell lines helped advance our knowledge about essential genes as well as synthetic lethality. Small screens targeting a subset of human genes in few cell lines contributed to the identification of synthetic dependencies in certain backgrounds (Bajrami et al., 2018; Berns et al., 2004; Dolly et al., 2017; Liu et al., 2012; Pathak et al., 2015). On a larger scale, the project DRIVE (McDonald et al., 2017) investigated nearly 400 cancer cell lines and 7,800 genes using a small interfering RNA (siRNA) essentiality screen. The screen found several synthetic lethal candidates including genes from parallel pathways (pathways performing similar functions) and vertical pathways (one pathway located downstream of the other). It also identified synthetic lethality between paralogs, meaning gene pairs sharing sequence similarity which originated through genome duplication and

functional divergence. In addition, over 16,000 human genes were targeted in a screen of 72 breast, pancreatic and ovarian cancer cell lines to characterize genetic interactions (Marcotte et al., 2012).

RNAi mediated knockdown results in off-target effects and variable efficiencies and is only transient (Birmingham et al., 2006; Evers et al., 2016; Jackson and Linsley, 2010). Besides these technical issues, cancer cell lines are a model system, do not reflect tumour environment and heterogeneity and may have adapted to different laboratory conditions (O'Neil *et al.*, 2017). These issues lead to unreproducible results between individual screens as highlighted by a study comparing 30 RNAi screens (Bhinder and Djaballah, 2013). This global analysis revealed that approximately 30% of the human genome plays an essential role in cell survival; however, 80% of short hairpin RNA (shRNA) hits and 88% of siRNA hits were only identified in one screen.

More recently, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) Cas9 screens have expanded our knowledge of gene essentiality and synthetic lethality. In contrast to transient RNAi, CRISPR Cas9 based gene editing directly affects the DNA and is therefore permanent (O'Neil *et al.*, 2017). A single guide RNA (gRNA) screen of 18,166 genes in two cell lines found that essential genes contain few inactivating mutations in the human population (Wang *et al.*, 2015). These genes are considered as core essential genes. In addition, a screen of 17,661 genes in five cell lines found that the essentiality of some genes depends on cell line specific vulnerabilities (Hart et al., 2015). These genes were named context dependent essential genes. Results were confirmed in a larger screen of over 200 cancer cell lines and 18,000 genes, providing a comprehensive resource of cancer vulnerabilities and potential new targets (Behan et al., 2019). A comprehensive resource collecting various RNAi and CRISPR Cas9 knockout screens is the Cancer Dependency Map (DepMap) project, currently containing essentiality data from over 2000 cancer cell line models (Tsherniak et al., 2017) (https://depmap.org/portal/).

Combinatorial CRISPR Cas9 screens are useful tools to identify synthetic lethal pairs. A screen targeting all combinations of 73 cancer genes in three cell lines identified 152 synthetic lethal interactions between gene pairs (Shen et al., 2017).

However, none of them were identified across all three cell lines. Context-dependency was also observed for the synthetic lethal interaction between *DUSP4*/*DUSP6* (Ito *et al.*, 2021). Their dual knockout reduced cell proliferation in *NRAS* mutant melanoma cell lines, whereas wild type cell lines remained unaffected. Indication of overlapping synthetic lethal interactions within the *EGFR*/*RAS*/*RIT1* signalling network were found by screening a lung cancer cell line for genetic interactions (Vichas et al., 2021). While 42% of genetic dependencies were shared between *EGFR*, *RAS* and *RIT1* mutants, 30% were specific to one alteration. Functional compensation through paralogs contributes to variable essentiality profiles of cancer cell lines (De Kegel and Ryan, 2019). Therefore, a recent CRISPR Cas9 knockout screen focused on the knockout of 2,060 human genes forming 1,030 paralog pairs in two cell lines (Parrish et al., 2021). Combining both cell lines, 12% of these paralogs exhibited synthetic lethal interaction, including potential novel drug targets for cancer. Combinations of drug exposure and CRISPR Cas9 knockout can directly identify clinically relevant synthetic lethal interactions. For example, a genome-wide screen identified the synthetic lethal interaction between *PLK1* loss and MEK inhibitors (Yu et al., 2021).

Similar to RNAi, CRISPR Cas9 knockout results in off-target effects (Alkan et al., 2018). In addition, silent 3-basepair insertions may reduce knockout efficiency, and genes in highly amplified regions of the genome are frequently scored as false positives (Munoz et al., 2016). These limitations can be addressed by optimizing the gRNA design and computational prediction of off-target effects (Akcakaya et al., 2018; Alkan *et al.*, 2018; Chakrabarti et al., 2019).

While siRNA knockdown is relatively time and cost efficient compared to CRIPSR Cas9 gene editing (Smith et al., 2017), CRISPR Cas9 knockout screens have several advantages over RNAi based screens. In contrast to the transient effect of RNAi knockdown, CRISPR Cas9 knockout directly alters the DNA and is therefore permanent. In addition, CRISPR Cas9 knockout is more efficient, consistent and induces less off-target effects (Evers *et al.*, 2016). The higher number of essential genes identified through CRISPR Cas9 screens implies that a full gene inactivation is necessary and CRISPR Cas9 screens are therefore

more accurate (Munoz *et al.*, 2016). Finally, a good alternative to the time-consuming cloning of Cas9 into cell lines is vector free gene editing through transfection of the Cas9 protein and gRNAs into the cell (Benedetti *et al.*, 2017).

1.3.3 Computational prediction of synthetic lethal interactions

Several computational prediction pipelines were developed to prioritize promising candidates for synthetic lethal interactions. Similar to experimental efforts, early predictions were done in *saccharomyces cerevisiae*. Synthetic sick or lethal interactions were identified using PPINs as input for support vector machines (Paladugu et al., 2008) as well as random walks (Chipman and Singh, 2009). To identify synthetic lethal gene pairs in humans, the DAta mIning SYnthetic lethality identification pipeline (DAISY) (Jerby-Arnon et al., 2014) used cancer alteration and expression data and identified gene pairs which were co-mutated less often than expected. In addition, it evaluated essentiality screens to identify genes that became essential when their synthetic lethal partners were under-expressed or mutated. Finally, expression data were used to identify synthetic lethal partners since they tend to be co-expressed. Similarly, the Mining Synthetic Lethals (MiSL) algorithm used pan-cancer copy number variation and RNA sequencing data to find potential synthetic lethal partners (Sinha et al., 2017). The Synthetic Lethal analysis via Network topology (SLant) approach predicted human synthetic lethal interactions through conserved protein interactions within and across species (Benstead-Hume et al., 2019). This approach performed best when integrating interaction data from different species. A recent study aimed to predict the most likely synthetic lethal interactors among paralog pairs (De Kegel *et al.*, 2021). By combining CRISPR Cas9 knockout data from over 700 cancer cell lines, they identified over 20 features of synthetic lethal interactors such as sequence conservation, shared protein interactions, participation in complexes and evolutionary conservation. This allowed a machine learning classifier to predict results of combinatorial CRISPR Cas9 screens and estimate the cell line specificity of synthetic lethal interactions.

Computational methods are an efficient and, compared to large scale screenings, cost-efficient way to narrow down the search for synthetic lethality. However, they are limited by incompleteness, and technical and biological variability in datasets, leading to variable outcomes. Whereas a joint analysis of several screens may be able to address technical reproducibility issues (Zamanighomi et al., 2019), a high degree of variability may also stem from biological variability. While synthetic lethal interactors in principle represent interesting targets for cancer therapy, a critical criterion for their translational value is the alteration of one of them in cancer to create a cancer-specific vulnerability.

1.3.4 Clinical relevance of synthetic lethality in cancer therapy

Originally, the term synthetic lethality described the negative genetic interaction between genes. It has now found application in the medical field and its meaning was expanded to include negative interactions between a certain genetic background and therapeutic treatment (Hartwell et al., 1997). Synthetic lethality is especially interesting as a new avenue to address intractable targets, such as lost tumour suppressors (Brunen and Bernards, 2017; Rehman *et al.*, 2010). Since the inhibition of genes is usually more straightforward and successful than the repair of a gene defect, addressing the loss of tumour suppressors through synthetic lethality-based therapies is a promising approach. In principle, it also leads to minimal side effects in non-cancer cells, since only the cancer cells have the alteration targeted by the therapy (McLornan et al., 2014) (Figure 1-3).

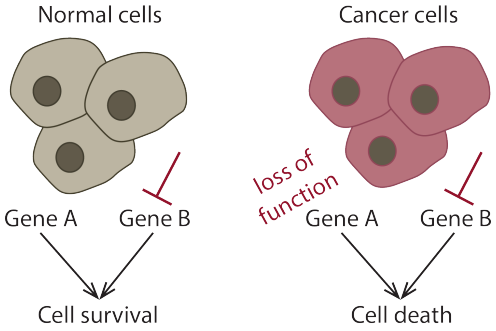


Figure 1-3 Synthetic lethality based on functional compensation between paralogs. Healthy cells can survive the loss of either Gene A or Gene B since A and B functionally compensate for each other. When a cancer cell loses the function of one gene through a loss-of-function mutation, the other gene becomes a target to induce cancer cell death.

Several approaches are of interest for cancer drug development based on synthetic lethality. First, targeting parallel pathways has proved successful, with the DNA damage repair pathway representing a very popular target (Das *et al.*, 2019; Li et al., 2020). Exploiting the negative genetic interaction between two DNA damage repair pathways, PARP inhibitors for treatment of *BRCA1* or *BRCA2* mutated breast cancers represent the first, and currently the only, approved cancer treatment based on synthetic lethality (Fong *et al.*, 2009; Fong *et al.*, 2010; Lord and Ashworth, 2017). Upon a successful clinical trial (Kaufman et al., 2015), the Food and Drug Administration approved the first PARP inhibitor olaparib in the clinic in 2014; niraparib and rucaparib were approved later (Food and Drug Administration, 2022). Several ongoing clinical trials are now investigating new generations of PARP inhibitors, combination treatments and applicability in several cancer types (ClinicalTrials, 2022).

Further potential synthetic-lethality-based therapeutic targets involving the DNA damage repair pathway are currently being investigated using cell lines. For example, *ATR* is essential in the repair of chromosome breaks during DNA replication (Brown and Baltimore, 2003), and its loss renders *TP53* essential to mediate DNA repair (Bieging et al., 2014; Kwok et al., 2016; Ruzankina et al., 2009). Synthetic dosage lethality may also be a promising approach. For instance, triple negative breast cancer cell lines overexpressing *MYC* are

especially sensitive to *PIM1* kinase inhibitors (Horiuchi et al., 2016). In addition, inhibition of the mitotic kinase *PLK1* leads to prometaphase accumulation and cell death in Ras mutant cells (Luo et al., 2009). Targeting exchangeable subunits that are key members of essential protein complexes can be successful. This was shown for *ARID1A* deficient ovarian cancer cell lines, which are vulnerable to *ARID1B* suppression based on the essential participation of the exchangeable subunits *ARID1A*/*ARID1B* in the SWI/SNF complex (Helming *et al.*, 2014).

Cancer cell line screens such as the DepMap project (Tsherniak *et al.*, 2017) provide the ideal basis for the identification of clinically relevant synthetic lethal interactions. For example, information from DepMap revealed that the loss of the *WRN* encoded RecQ DNA helicase is incompatible with microsatellite instability in cancer cell lines. DepMap was also used to discover that *NXF1* is rendered essential in *MYCN*-amplified neuroblastoma (Malone et al., 2021). In an independent CRISPR Cas9 double knockout screen, synthetic lethal interactions between 87 tumour suppressors and their paralogs were identified in a lung adenocarcinoma cell line (Parrish *et al.*, 2021). Not only synthetic lethal partners of lost tumour suppressors directly may be clinically relevant, but also partners of passengers frequently co-deleted with tumour suppressors. Lord et al. identified a higher sensitivity towards *DDX5* loss in cell lines with deletions of *DDX17*, a gene in close proximity to and frequently co-deleted with the tumour suppressor *DYH9* (Lord et al., 2020). Similarly, *VPS4B* is often co-deleted with the tumour suppressor *SMAD4*, rendering its synthetic lethal partner *VPS4A* a potential therapeutic target (Lord *et al.*, 2020; Neggers et al., 2020).

To select tumour suppressor genes recurrently lost in cancer and therefore especially interesting for the development of new synthetic lethality-based therapies, TCGA offers a rich resource of cancer genomics data (TCGA Research Network, 2021). In a comprehensive analysis of 9,423 tumour exomes from TCGA, Bailey et al. identified 299 driver genes damaged by somatic mutations and their enrichment across cancer types (Bailey *et al.*, 2018). Among these, the most frequent alterations in tumour suppressors represent interesting targets. In addition to somatic mutations, epigenetic silencing of tumour suppressors plays an important role, especially early in tumour initiation

(Kazanets *et al.*, 2016). For example, silencing of CDKN2A in mammary tissue leads to increased risk for breast cancer (Gauthier et al., 2007). Epigenetic silencing of tumour suppressors also plays an important role in many other cancer types such as ovarian cancer (Wrzeszczynski et al., 2011), acute myeloid leukaemia (Cancer Genome Atlas Research et al., 2013), head and neck cancer (Kaur et al., 2010) and lung cancer (Hoang and Landi, 2022).

These are only few examples of synthetic lethality-based therapy approaches currently investigated in cell lines, animal models and clinical trials. However, designing targeted therapies based on synthetic lethality is not straightforward. Modulating factors such as the mutational background or tumour environment may affect genetic interactions (O'Neil *et al.*, 2017). For example, loss of *53BP1* can suppress the synthetic lethal interaction between PARP inhibitors and *BRCA* mutations through promotion of *ATM* dependent DNA repair (Bunting et al., 2010; Jaspers et al., 2013). Introducing hypoxic conditions in cell culture increases the sensitivity of cancer cells towards PARP inhibitors (Chan et al., 2010) and changes in media conditions lead to identification of different synthetic lethal interactors (Ku et al., 2020). These results indicate that several synthetic lethal interactions may be missed due to experimental conditions being unrepresentative of the tumour environment *in vivo*. In addition, they highlight the challenges in reproducibility and translation from cell lines to the clinic. The heterogeneity of synthetic lethal interactions was confirmed by comparing three different studies investigating synthetic lethal interactions with *KRAS* mutations (Ku *et al.*, 2020). While interactions differed substantially on a gene level, they occurred within the same pathways. Synthetic lethal interactors of both *KRAS* and *MYC* alterations which were identified through different studies were significantly more likely to directly interact with each other or participate in the same protein complex (Ku *et al.*, 2020). Therefore, a network analysis approach may be more robust than single gene screens.

The mutual exclusivity of altered expression or alterations between gene pairs in cancer samples indicates a disadvantage for the cancer (Szczurek et al., 2013). Thus, Szczurek et al. used somatic mutations, copy number variations and expression values in glioblastoma samples to identify clinically relevant targets.

A similar approach was further pursued by the Identification of clinically relevant Synthetic LEthality (ISLE) algorithm. It filtered 16 million synthetic lethal interactions identified from large scale cell line screens for underrepresentation of co-inactivation in TCGA (Lee et al., 2018). Gene pairs were prioritized if their co-inactivation led to better patient survival and if they showed phylogenetic evolutionary proximity. Together, these criteria predicted patient response to drug treatment. These methods place a clear focus on the clinical parameters, while biological parameters such as paralogy or protein interactions are not considered.

1.4 Epigenetics and cancer

1.4.1 Defining epigenetics

DNA alterations are crucial for cancer cells to induce diversity and thus adapt to external selective pressures. In contrast to permanent alterations in the DNA, epigenetics is defined as 'the study of changes in gene function that are heritable and that do not entail a change in DNA sequence' (Wu and Morris, 2001).
These changes enable functional specialization of cells in multicellular organisms despite their identical genetic makeup. Epigenetic silencing also plays an essential role in X chromosome inactivation in female mammals (Fang et al., 2019). In addition, it allows for adaptation to environmental changes through reversible activation or downregulation of certain genes. This can confer evolutionary advantages both on species level and cellular level (Jaenisch and Bird, 2003).
Epigenetic changes are conferred by epigenetic modifiers and result in DNA modifications, post-translational modifications of histones, chromatin remodelling and RNA-based regulation (Feinberg et al., 2016; Gibney and Nolan, 2010; Plass et al., 2013). While non-coding RNAs play an important role in regulating transcript activities, sequence-specific recognition and catalysis (Gibney and Nolan, 2010), the following section will focus on proteins modifying epigenetic marks (Figure 1-4).

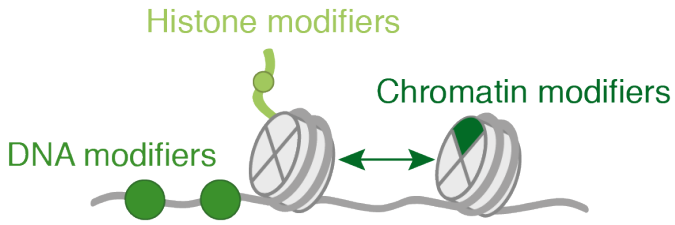


Figure 1-4 Three groups of epigenetic modifiers.
Proteins modifying epigenetic marks are involved in histone modification, DNA modification, chromatin modification, or a combination thereof.

DNA methylation is achieved through methylation of cytosines to form 5-methylcytosines (Siegfried and Cedar, 1997). Methylation mainly occurs in a CpG context, meaning cytosines neighbouring guanines are methylated. In general, DNA methylation is associated with condensed heterochromatin and silenced gene expression (Bird, 1992; Keshet et al., 1986). The proper establishment of methylation patterns is a requirement for successful embryonic development. Established patterns are inherited through mitosis and play a crucial role in gene silencing, resulting in tissue specific expression, differentiation (Smith and Meissner, 2013) and X chromosome inactivation (Jaenisch et al., 1998). Silencing of several developmental genes is mediated through methylation of promoter regions, which are especially rich in CpG stretches called CpG islands (Illingworth and Bird, 2009). Gene expression also negatively correlates with methylation of the first intron, a relationship conserved across vertebrates (Anastasiadi et al., 2018). In addition, gene inactivation through methylation is important for silencing of endogenous human retroelements (Schulz et al., 2006). In contrast to the methylation of regulating elements, methylation of the gene body has been shown to correlate with high gene expression in proliferating cells (Aran et al., 2011). DNA demethylation is achieved through two parallel routes. While active DNA demethylation is a multi-step oxidation process, passive demethylation occurs through dilution throughout several cell divisions (Wu and Zhang, 2017). Both play an important role in the epigenetic reprogramming process during pre-implantation development.

The first level of DNA packaging is achieved through the formation of the nucleosome. It consists of a histone octamer core, formed by two molecules of

each subunit H2A, H2B, H3 and H4, and 147bp of DNA wrapped around it (Alberts et al., 2015). The linker histone H1 binds DNA between nucleosomes. While DNA is tightly wrapped around the core of histones, flexible protruding histone tails are accessible for various post-translational modifications (Kouzarides, 2007). Histone modifications can have a repressive or activating impact on gene transcription (Zhao and Shilatifard, 2019). For instance, histone 3 lysine 9 trimethylation (H3K9me3) and histone 3 lysine 27 trimethylation (H3K27me3) modifications are typically present in inactive chromatin states. In contrast, acetylation marks such as histone 4 lysine 16 acetylation (H4K16ac) or histone 3 lysine 27 acetylation (H3K27ac) are prominent histone modifications indicating active chromatin states. Further histone marks include ubiquitination, biotinylation, sumoylation, citrullination, crotonylation, buturylation and propionylation. In contrast to DNA methylation, histone modifications are highly dynamic and allow flexible adaptation of transcription levels (Zentner and Henikoff, 2013). Histone acetylation marks have half lives in the order of minutes, whereas methylation marks are more stable with half lives in the order of days (Barth and Imhof, 2010).

Chromatin remodellers change the three-dimensional structure of chromatin (Hargreaves and Crabtree, 2011). They frequently engage in complexes containing an adenosine triphosphatase (ATPase) subunit. It supplies energy to replace histone subunits, remove nucleosomes or change nucleosome positions and thus their distance from each other (Saha et al., 2006). This common ATPase unit is combined with differing subunits conferring a unique specificity to each complex. A prominent member of the chromatin remodellers is the SWI/SNF complex which re-arranges nucleosomes to form a disordered array (Owen-Hughes et al., 1996). It opens up long stretches of accessible DNA by pushing the flanking nucleosomes closer together. In contrast, the Imitation SWItch (ISWI) complex family promotes equal spacing between nucleosomes (Varga-Weisz et al., 1997). Other complexes are involved in chromatin conformation changes without relying on adenosine triphosphate (ATP) hydrolysis. For example, the mediator complex is essential for DNA looping to allow successful RNA transcription (Soutourina, 2018).

Some modifications also have an indirect impact on others. For example, histone acetylation leads to an open chromatin conformation (Grunstein, 1997), whereas DNA methylation leads to an inaccessible, compact structure (Keshet *et al.*, 1986). In addition, DNA methylation interacts with different histone marks (Jeziorska et al., 2017; Newell-Price et al., 2000).

The classification of proteins as being involved in epigenetics depends on the source. For example, a study by Plass et al. originally collected a list of 709 genes connected to epigenetic modification (Plass *et al.*, 2013) which was then expanded to over 1000 genes. The database Epifactors collected 815 genes involved in epigenetics (Medvedeva et al., 2015). Less inclusive lists include 263 genes in a study by Hoffman et al. (Hoffman *et al.*, 2014) and a list of 212 genes provided by Yang et al. (Yang et al., 2015).

1.4.2 Epigenetic modifiers and cancer

Epigenetic modifiers are frequently mutated in cancer (Feinberg *et al.*, 2016). Over 70% of 709 epigenetic modifiers are mutated in at least two ICGC samples (Plass *et al.*, 2013) and over 50% of cancers harbour mutations in genes involved in chromatin organization (Jones et al., 2016). A possible explanation for this high frequency is the broad impact of epigenetic modifiers on transcriptional regulation: The deregulation of a single epigenetic modifier can affect expression states of many target genes (Feinberg *et al.*, 2016).

Epigenetic deregulation in human cancer tissues was first described in 1983, when it was discovered that cancerous lung and colon tissues were hypomethylated (Feinberg and Vogelstein, 1983). Since then, deregulation of all three classes of epigenetic modifiers (DNA modifiers, chromatin modifiers and histone modifiers) has been described (Han et al., 2019; Plass *et al.*, 2013; Zhao and Shilatifard, 2019). Approximately 20% of acute myeloid leukaemia cases harbour a mutation in the DNA methyltransferase *DNMT3A*, resulting in enhanced proliferation and shorter overall survival (Ley et al., 2010; Xu et al., 2014). On the other hand, mutations in the *TET2* gene belonging to the DNA demethylation pathway induce myeloid tumours (Ko et al., 2010). Aberrant

histone modifications also drive cancer formation. Damaging *CREBBP* alterations occur frequently in B-cell carcinoma (Jiang et al., 2017) as well as small cell lung cancer (Jia et al., 2018), resulting in depletion of H3K27ac. As a result, these cancers respond well to histone deacetylase *HDAC3* inhibition. A prominent example of histone methylation deregulation is the duplication or translocation of the histone methyltransferase *MLL* (Krivtsov and Armstrong, 2007). Loss of *MLL* through translocation leads to histone 3 lysine 4 (H3K4) methylation loss and can have further effects on epigenetics depending on the translocation partner. For example, the inactivation of histone methyltransferase *DOT1L* leads to reduction of histone 3 lysine 79 (H3K79) methylation. An example of aberrant chromatin organization is the alteration of *SMARCA4* and several other subunits of the chromatin remodelling SWI/SNF complex in medulloblastoma (Jones et al., 2012; Parsons et al., 2011; Pugh et al., 2012).

Given the high frequency of epigenetic modifier alterations in cancer, they represent valuable targets for therapy (Pfister and Ashworth, 2017). For example, DNA methyltransferase inhibitors are used as cancer therapy to counteract the hypermethylation and thus silencing of tumour suppressors (Gnyszka, Jastrzębski and Flis, 2013). Histone deacetylase inhibitors are used in cancers over-expressing histone deacetylases, resulting in less condensed chromatin (Ceccacci and Minucci, 2016). The histone methyltransferase inhibitor tazemetostat was approved in 2020 by the Food and Drug Administration (FDA) to treat patients with locally advanced or metastatic epithelioid sarcoma (Food and Drug Administration, 2020). However, drugs targeting epigenetic modifiers may non-specifically interact with proteins not involved in epigenetics. Thus, it is not surprising that side effects can be severe (Pfister and Ashworth, 2017). In addition, many epigenetic modifiers are tumour suppressors (Cohen et al., 2011; Esteller, 2006; Wilson and Roberts, 2011). Cancer therapy has traditionally relied heavily on the inhibition of oncogenes, as drug development is more straight forward (Morris and Chan, 2015). Therefore, new therapy approaches to target epigenetic tumour suppressors are needed.

1.4.3 Synthetic lethality in epigenetic modifiers

Synthetic lethal interactions between paralogs have been observed between several epigenetic modifier pairs. For example, the two pairs *ARID1A*/*ARID1B* and *SMARCA2*/*SMARCA4* are interchangeable subunits of the SWI/SNF complex (Mashtalir *et al.*, 2018) and synthetic lethal partners (D'Antonio *et al.*, 2013; Helming *et al.*, 2014; Hoffman *et al.*, 2014; Oike *et al.*, 2013). In addition, the cohesin subunits *STAG1*/*STAG2* can compensate for each other's loss (Benedetti *et al.*, 2017; van der Lelij *et al.*, 2017). Furthermore, synthetic lethal interactions between histone methyltransferases *EZH1*/*EZH2* (Honma *et al.*, 2017) and *KMT2A*/*KMT2B* (Ernst *et al.*, 2016) as well as histone acetyltransferases *CREBBP*/*EP300* (Ogiwara *et al.*, 2016) were observed.
Synthetic lethality between parallel epigenetic pathways was also observed. *EZH2* inhibition is lethal in combination with damaging mutations of members of the SWI/SNF complex, *ARID1A* (Bitler et al., 2015) or *SMARCB1* (Knutson et al., 2013). In addition, cell lines deficient in histone 3 lysine 36 trimethylation (H3K36me3) marks, for example caused through *SETD2* mutations, are hypersensitive to inhibition of the kinase *WEE1* (Pfister et al., 2015). Finally, besides their synthetic lethal interaction with *BRCA1* and *BRCA2* mutations, PARP inhibitors also represent a highly effective cancer treatment in combination with low *KMT2C* expression (Rampias et al., 2019). This is based on the downregulation of DNA damage response genes upon downregulation of *KMT2C*.

1.5 Aim of the thesis

This thesis aims to expand the knowledge about cancer driver genes and place their properties into context by characterizing diverse gene groups in health and disease. It then aims to identify potential cancer vulnerabilities based on functional compensators using a subset of these properties.
Firstly, the contributions to NCG from this thesis are discussed. NCG is a collaborative group project collecting and characterising genes implicated in

cancer and lays the foundations for the following two results chapters. Previous versions of NCG include a comprehensive collection of genes driving cancer upon alterations in their coding sequence. Recent screens have identified cancer driving alterations in non-coding regions of the genome, as well as alterations driving clonal expansion in non-cancer tissues. Aiming to understand the differences between drivers of clonal expansion in different contexts, NCG introduces two new driver gene categories: non-coding drivers and healthy drivers. In addition, evolutionary properties of cancer genes investigated by NCG have so far included gene duplicability, evolutionary origin, miRNA interactions, RNA and protein expression as well as position in the PPIN and participation in protein complexes by the encoded protein. Beyond updating these properties, the most recent version (NCG7) adds gene essentiality and tolerance to germline variation to the property profile. These additions contribute to a more refined characterization of different groups of driver genes.

The second results chapter aims to understand the properties of cancer driver genes in the wider context of genes involved in health and disease. So far, relationships between several evolutionary properties and gene groups such as essential genes, cancer genes or Mendelian disease genes have been investigated individually. To provide a more comprehensive picture, we combine nine evolutionary properties using a principal component analysis and a random forest classifier. This integration allows us to compare evolutionary properties of genes involved in 25 biological pathways and their tolerance to germline and somatic loss-of-function alterations as well as involvement in genetic diseases. We then apply our comprehensive understanding of evolutionary properties to prioritize cancer driver candidates for functional validation.

The third results chapter focuses on implementing previously gained knowledge to identify synthetic lethal interactions between paralogs. In specific, we predict synthetic lethal pairs based on gene duplicability, protein complex formation, dependency of essentiality on mutation and expression as well as alteration frequency in cancer. This allows us to identify synthetic lethal interactions based on a subset of previously annotated properties and also prioritize clinical relevance. We experimentally validate predictions in cancer cell lines using

siRNA and CRIPSR Cas9 based knockout in cancer cell lines. This results in the identification of two new synthetic lethal paralog pairs, one of which is context dependent. Given their high alteration frequency in cancer, context dependent essentiality and previous observation of their synthetic lethal interactions, we investigate interactions between epigenetic modifiers, aiming to identify evolutionary properties which may predispose epigenetic modifiers to synthetic lethal interactions.

Chapter 2. Materials & Methods

2.1 The Network of Cancer Genes

2.1.1 NCG6

The extraction of 2372 cancer genes, including 711 canonical cancer genes, was based on two sources of established cancer genes (Futreal *et al.*, 2004; Vogelstein et al., 2013) and 273 further cancer sequencing screens (Repana *et al.*, 2019). Each source was reviewed by at least two independent scientists.
Applying previously published methods, 19,549 unique gene loci were identified (Rambaldi *et al.*, 2008). Briefly, protein sequences were obtained from RefSeq v.85 (O'Leary et al., 2016) and were aligned to the human genome (hg38) using the basic-local-alignment-search-tool-like alignment tool (BLAT) (Kent, 2002). Unique genomic loci were identified based on gene coverage, span and identity. A locus was considered a duplicate if it matched 60% of the protein sequence (i.e. 60% coverage). Interactions between human proteins were obtained and integrated from four databases (BioGRID v.3.4.157 (Chatr-Aryamontri et al., 2017), MIntAct v.4.2.10 (Orchard et al., 2014), DIP (02/2018) (Salwinski et al., 2004), HPRD v.9 (Keshava Prasad et al., 2009)). Only interactions with at least one original supporting source were considered to derive a human PPIN including 16,322 proteins and 289,368 interactions. The degree, betweenness and clustering coefficient of the PPIN were calculated using the igraph R package v.1.2.6 (Csardi and Nepusz, 2005). Data on human protein complex formation were integrated from three sources (CORUM (07/2017) (Ruepp et al., 2010), HPRD v.9 (Keshava Prasad *et al.*, 2009), Reactome v.63 (Fabregat et al., 2018)) for 8,080 human proteins. Only interactions with at least one original supporting source were considered. Gene essentiality data were obtained from two sources including shRNA knockdown and CRISPR Cas9 knockout screens in human cell lines (PICKLES (09/2017) (Lenoir et al., 2018), OGEE v.2 (Chen et al., 2017)). Annotations of gene essentiality were retained as in the original database (PICKLES: genes with a Bayes Factor >3 were essential, OGEE: genes were annotated as “essential” or “nonessential). The nonredundant union of both

databases resulted in essentiality information on 18,833 genes in 178 cell lines. Orthologous genes were obtained from EggNOG v.4.5.1 (Huerta-Cepas et al., 2016) and used to identify the gene evolutionary origin of 18,486 human genes as described before (D'Antonio and Ciccarelli, 2011). RNA sequencing data of healthy human tissues were combined from two sources (Protein Atlas v.18 (Uhlen et al., 2015), GTEx v.7 (Consortium, 2013)) for 18,984 genes. Expression values were obtained for 37 tissues from Protein Atlas and genes were considered as expressed at expression values higher than 1 transcript per million (TPM). Expression values for 11,688 samples including 30 tissue types were obtained from GTEx and genes were considered as expressed if their median expression value across the respective tissue type was greater than 1TPM. RNA expression of 1,561 cancer cell lines was obtained from three sources (Cancer Cell Line Encyclopedia (02/2018) (Cancer Cell Line Encyclopedia and Genomics of Drug Sensitivity in Cancer, 2015), COSMIC Cancer Cell Line Project (v.84) (Futreal *et al.*, 2004), Genentech study (06/2014) (Klijn et al., 2015)). Genes were considered as expressed in the Cancer Cell Line Encyclopedia and the Genentech study if their expression values were greater than 1 read per kilobase million (RPKM) and were annotated as over, under or normally expressed as determined by the COSMIC Cancer Cell Line Project. Protein expression from immunohistochemistry assays of healthy human tissues for 13,001 proteins was obtained from Protein Atlas v.18 (Uhlen *et al.*, 2015). Detection levels were reported as in the original source as not detected, low, medium, or high expression in 44 tissues, and the highest reported value was retained for multiple cell types of the same tissue. Interactions between human genes and miRNAs that were confirmed by experimental validation were obtained from miRTarBase v.7.0 (Chou et al., 2018) and miRecords v.4.0 (Xiao et al., 2009) and integrated to form an interaction network of 14,649 genes and 1,762 miRNAs. The degree and betweenness of the miRNA interaction network were calculated using the igraph R package v.1.2.6 (Csardi and Nepusz, 2005). Annotation of functional pathways was obtained from Reactome v.63 (Fabregat *et al.*, 2018) and KEGG v.85.1 (Kanehisa et al., 2017) for 11,344 human proteins. A two-sided Wilcoxon test was used to compare distributions of protein and miRNA network properties,

the number of cell lines in which genes were essential and tissues in which genes or proteins were expressed. A two-sided Fisher's exact test was used to compare proportions of duplicated, pre-metazoan, essential genes and proteins engaging in complexes.

2.1.2 NCG7

By conducting a literature search for cancer sequencing screens published up to 2020, 37 new cancer screens were added to the previous list of 273 screens, reaching a total of 310 cancer screens (Dressler *et al.*, 2021). This included 19 screens identifying cancer drivers altered in non-coding regions of the genome. Further, 18 screens regarding sequencing of non-cancer tissues were included. Each source was reviewed by at least two independent scientists. Canonical cancer drivers were combined from three sources ((Saito et al., 2020; Vogelstein *et al.*, 2013), Cancer Gene Census v.91 (Tate *et al.*, 2019)). All Tier 1 and Tier 2 genes extracted from the Cancer Gene Census were retained except for drivers identified by gene fusions. Genes were further annotated as tumour suppressors, oncogenes or drivers with dual role based on the consensus annotation of the original sources. Identified drivers which were among 148 potentially false positive genes (Lawrence *et al.*, 2013; Saito *et al.*, 2020) were only included if they passed further manual checks of supporting evidence. Genes were annotated as canonical cancer drivers, candidate cancer drivers (identified from cancer sequencing screens only), drivers with alterations in coding or non-coding regions and healthy drivers (identified from sequencing screens in non-cancer tissues).

Unique genomic loci and gene duplicates were identified for 19,756 genes as described in Chapter 2.1.1, with the alignment score considered in addition to gene coverage, span and identity to identify best hits of the alignment. Five sources (BioGRID v.3.5.185 (Oughtred et al., 2019), IntAct v.4.2.14 (Orchard *et al.*, 2014), DIP (02/2018) (Salwinski *et al.*, 2004), HPRD v.9 (Keshava Prasad *et al.*, 2009) and Bioplex v.3.0 (Huttlin et al., 2021)) were integrated to obtain a total of 542,397 binary interactions between 17,883 proteins. The degree,

betweenness and clustering coefficient of the PPIN were calculated using the igraph R package v.1.2.6 (Csardi and Nepusz, 2005). Protein complex interactions were obtained for 8,504 proteins and 9,476 complexes by integrating three sources (CORUM v.3.0 (Giurgiu et al., 2019), HPRD v.9 (Keshava Prasad *et al.*, 2009) and Reactome v.72 (Jassal et al., 2020)). Three RNAi knockdown and six CRISPR Cas9 knockout screens (Behan *et al.*, 2019; Dempster et al., 2019; DepMap Broad, 2019a; b; 2020; Lenoir *et al.*, 2018; McFarland et al., 2018; Meyers et al., 2017; Tsherniak *et al.*, 2017) were integrated to obtain essentiality data for 19,013 genes and 1,122 cell lines. Retaining the original definitions of essentiality, genes with a CERES (Meyers *et al.*, 2017) or DEMETER (Tsherniak *et al.*, 2017) score <-1 or Bayes score (Hart and Moffat, 2016) >5 were considered as essential. The evolutionary origin of 18,922 human genes was derived using ortholog annotation from EggNOG v.5.0 (Huerta-Cepas et al., 2019). The union of Protein Atlas v.19.3 (Uhlen *et al.*, 2015) and GTEx v.8 (Consortium, 2020) provided RNA expression of 19,231 genes in 49 healthy tissues and genes were considered as expressed at expression values >1TPM. Protein expression values for 13,229 proteins in 45 healthy tissues were obtained from Protein Atlas v.19.3 (Uhlen *et al.*, 2015), and the highest expression value was considered if several values per tissue were available. Experimentally supported interactions from miRTarBase v.8.0 (Huang et al., 2020) and miRecords v.4.0 (Xiao *et al.*, 2009) were integrated to obtain interactions between 14,747 genes and 1,758 miRNAs. The degree and betweenness of the miRNA interaction network were calculated using the igraph R package v.1.2.6 (Csardi and Nepusz, 2005). gnomAD v.2.1.1 (Karczewski et al., 2020) was used to obtain the loss-of-function observed/expected upper bound fraction (LOEUF) score for 18,392 genes. Germline alterations (single nucleotide variants (SNVs) and indels) were derived from the combination of 2,504 samples from the 1000 Genomes Project Phase 3 v.5a (Genomes Project et al., 2015) and 125,748 samples from gnomAD v.2.1.1 (Karczewski *et al.*, 2020). ANNOVAR (October 2019) (Wang et al., 2010) was used to annotate the damaging effect of genes and the following mutations were defined as damaging:

1) all truncating (stop gain, stoploss, frameshift) mutations,
2) missense mutations predicted as damaging by at least five of seven function-based predictions and two of three conservation-based predictions (Table 2-1),
3) splicing mutations predicted as damaging by at least one of two ensemble scores (Table 2-1).

Table 2-1 Prediction tools used to assess damaging effect of mutations.
Different combinations of function-, conservation- and splicing-based prediction tools were used to identify damaging missense mutations.

Prediction tool	Method	Reference
SIFT	function	(Kumar et al., 2009)
PolyPhen-2 HDIV	function	(Adzhubei et al., 2013)
PolyPhen-2 HVAR	function	(Adzhubei *et al.*, 2013)
MutationTaster	function	(Schwarz et al., 2010)
MutationAssessor	function	(Reva et al., 2011)
LTR	function	(Chun and Fay, 2009)
FATHMM	function	(Shihab et al., 2013)
PhyloP	conservation	(Pollard et al., 2010)
GERP++RS	conservation	(Davydov et al., 2010)
SiPhy	conservation	(Garber et al., 2009)
ADA	splice site	(Liu et al., 2016)
RF	splice site	(Liu *et al.*, 2016)

gnomAD v.2.1.1 (Karczewski *et al.*, 2020) was used to derive 32,558 germline structural variants from 15,708 samples. The numbers of damaging mutations and structural variations per base pair were calculated for each gene.

A two-sided Fisher's exact test was used to compare proportions of duplicated, pre-metazoan, essential genes and proteins engaging in complexes. A two-sided Wilcoxon test was used to compare distributions of tissues in which genes or proteins were expressed, protein and miRNA network properties, LOEUF scores,

damaging mutations and structural variants per base pair. Benjamini-Hochberg multiple comparison correction within each property was performed.

To display properties in a heatmap, a normalized property score was calculated for each evolutionary property and each driver group d as

$$Normalized\ property\ score = sgn(\Delta_d) \times \frac{|\Delta_d| - min_t|\Delta_t|}{max_t|\Delta_t| - min_t|\Delta_t|} = \frac{\Delta_d}{max_t|\Delta_t|}$$

where t represents ten gene groups (canonical drivers, candidate drivers, tumour suppressors, oncogenes, drivers with coding alterations, drivers with noncoding alterations, canonical healthy drivers, candidate healthy drivers, remaining healthy drivers and rest of human genes); Δ_d represents the difference of medians (continuous properties) or proportions (categorical properties) between each driver group and the rest of human genes; and $sgn(\Delta_d)$ is the sign of the difference. Minima and maxima were calculated across all eleven gene groups for each property.

The PPIN display was produced with the igraph R package (Csardi and Nepusz, 2005) and integrated with the database and website using the shiny (Chang et al., 2021), shinyWidgets (Perrier et al., 2021), RMySQL (Ooms et al., 2021) and DBI (Wickham et al., 2021) packages.

2.2 Integration of evolutionary properties

2.2.1 Annotation and integration of gene evolutionary properties and function

Gene evolutionary origin, duplicability, breadth of mRNA and protein expression, miRNA-gene interactions, engagement in protein complexes and protein-protein interactions as well as participation in Reactome and KEGG functional pathways were annotated for subsets of 19,549 genes and their encoded proteins as described for NCG6 in Chapter 2.1.1 (Table 4-1, Table 7-1).

Evolutionary properties were integrated into the evolutionary property score (EP score) using a random forest approach. The positive training set consisted of 148 genes, including the intersection of genes with property values in the respective top quartile for each of the seven numerical evolutionary properties (categorical

properties were not considered in this step). The negative training set consisted of 117 genes, obtained through the intersection of genes in the respective bottom quartile. All nine evolutionary properties were included in the training of the random forest classifier as implemented in the randomForest package in R (Liaw and Wiener, 2002). The predicted likelihood of a gene belonging to the positive training set was then defined as the EP score, therefore indicating high numerical property values.

The principal component analysis (PCA) was conducted and visualized using R packages PCAtools (Blighe and Lun, 2020), stats (R Core Team, 2020) and ggbiplot (Vu, 2011). Forty-three genes with a PPIN degree >500 dominated the PCA results and were therefore excluded. Values for all nine evolutionary properties were available for 4,306 genes, missing values were imputed using the missMDA package (Josse and Husson, 2016).

2.2.2 Germline and somatic alterations in human samples

The LOEUF score was obtained for 18,322 genes as described in Chapter 2.1.2. Somatic alterations (quality-controlled SNVs and indels, copy number alterations) and gene expression data for 7,953 samples including 34 cancer types were obtained from the NCI Genomic Data Commons (Grossman et al., 2016). Of these, 7,921 samples had at least one damaging alteration and 7,626 samples were retained after excluding hypermutated samples (Table 7-2). Copy number variant segments, sample ploidy and sample purity were obtained from TCGA single nucleotide polymorphism arrays using ASCAT v.2.5.2 (Van Loo et al., 2010). Genes were considered deleted if at least 25% of the gene were contained in the copy number aberrant region. Since ASCAT returns integer copy numbers, homozygous gene losses had a copy number equal to 0 and were filtered for RNA expression values < 1 Fragment Per Kilobase per Million (FPKM) over sample purity. Mutations were annotated with ANNOVAR (04/2018) (Wang *et al.*, 2010) and dbNSFP v.3.0 (Liu *et al.*, 2016) and only exonic or splicing mutations were retained. In line with the definition for damaging germline mutations (Chapter 2.1.2), the following somatic mutations were defined as damaging:

1) all truncating (stop gain, stoploss, frameshift) mutations,
2) missense mutations predicted as damaging by at least five of seven function-based predictions and two of three conservation-based predictions (Table 2-1),
3) splicing mutations predicted as damaging by at least one of two ensemble scores (Table 2-1).

Homozygous deletions and damaging mutations were considered as loss-of-function alterations.

2.2.3 Loss-of-function alterations and essentiality in cancer cell lines

Gene essentiality data were obtained as described in Chapter 2.1.2. A gene was considered essential in a cell line if at least one dataset declared it as essential. Information obtained from CRISPR knockout screens was prioritized due to higher reliability (Hart *et al.*, 2014). Genes were assigned to one of three categories:

1) Core essential: essential in at least 80% of tested cell lines
2) Context dependent essential: essential in more than one cell line and less than 80% of tested cell lines
3) Nonessential: essential in 0-1 cell line

Copy number alterations, mutations, RNA expression and essentiality for 18,511 genes in 878 cell lines were obtained from DepMap (DepMap Broad, 2020; Ghandi et al., 2019). A copy number of less than 0.25*cell line ploidy (equivalent to less than 0.5 in a diploid cell) was considered as a homozygous deletion if RNA expression was less than 1TPM. Loss-of-function mutations in cell lines were annotated and defined as described for somatic alterations in cancer samples in Chapter 2.2.2.

2.2.4 Disease genes

A total of 711 canonical cancer drivers, including 239 tumour suppressors and 239 oncogenes were obtained from NCG7 (Chapter 2.1.1)(Repana *et al.*, 2019).

In addition, 79 healthy drivers were obtained from a literature review (Wijewardhane *et al.*, 2020), 43 of which were also canonical cancer drivers. Of note, this was a subset of the 95 healthy drivers identified for NCG7 (Chapter 2.1.2). The OMIM database (7th September 2020) (McKusick-Nathans Institute of Genetic Medicine, 2020) was used to obtain 3,568 genes related to Mendelian Disease Disorders, including 1,982 genes associated with a recessive and 1,022 with a dominant phenotype and excluding genes annotated as “nondiseases”, susceptibility to infection or provisional classifications.

2.3 Prediction of synthetic lethal interactions

2.3.1 First prediction method

The first prediction method was based on a set of 19,014 human genes (An *et al.*, 2016). Duplicates were defined as additional alignments of the protein sequence to the human reference genome hg18 with at least 5% coverage. The annotation of copy number alterations and mutations in 7,828 tumour samples including 31 tumour types (Table 7-2) for 19,014 genes was performed based on data obtained from TCGA (Grossman *et al.*, 2016; TCGA Research Network, 2021). Sample purity was obtained from NCI Genomic Data Commons (Grossman *et al.*, 2016). Hypermutated samples, sequencing errors, technical biases, mutations with a variant allele frequency <10% and indels longer than five nucleotides were removed. Non-silent mutations were identified using ANNOVAR (Wang, Li and Hakonarson, 2010). Damaging mutations were predicted using function-, conservation- and splice site-based prediction tools. The following mutations were defined as damaging:

1) all truncating (stop gain, stoploss, frameshift) mutations,
2) missense mutations predicted as damaging by at least five of seven function-based predictions (Table 2-1),
3) missense mutations predicted as damaging by at least two of three conservation-based predictions plus one of seven function-based predictions (Table 2-1) or

4) splicing mutations predicted as damaging by at least one of two ensemble scores (Table 2-1).

Gene copy number was calculated from the segment mean obtained from TCGA level 3 profiles as $CN = 2^{segment\ mean} \times 2$ and rounded to the nearest integer. Genes were considered deleted if at least 25% of the gene were contained in the copy number aberrant region. Genes with a copy number <0.5 were considered as homozygously deleted, genes with 0.5< copy number <1.5 were considered as heterozygously deleted.

Homozygous deletions and double hits (heterozygous deletions with an additional damaging mutation on the other allele) were considered as loss-of-function alterations. Given that damaging mutations in combination with a heterozygous deletion should be observed in all sequenced alleles, double hits with a low allele frequency of the damaging mutation were discarded. To account for contaminating healthy tissues and mutation clonality, the allele frequency threshold was calculated for each sample based on its purity p and a minimal clonality of 75%. Therefore, the fraction of cells with a loss-of-function alteration was calculated as $0.75p$ and the fraction of wild type cells with 2 copies of the wild type allele as $1 - 0.75p$. Consequently, the threshold for the allele frequency AF of the damaging mutation in a double hit scenario was filtered to be $AF > \frac{0.75p}{0.75p+2\times(1-0.75p)}$.

2.3.2 Improved prediction method

The improved prediction method was based on a set of 19,756 human genes identified in NCG7 (Chapter 2.1.2) (Dressler *et al.*, 2021; Rambaldi *et al.*, 2008). Duplicates were defined as additional alignments of the protein sequence to the human reference genome with at least 20% coverage. Protein complex interactions were obtained as described in Chapter 2.1.2. Gene essentiality, copy number alterations, mutations and RNA expression in cell lines were obtained as described in Chapter 2.2.3. Essentiality dependency was defined as essentiality of gene A depending on RNA expression or alteration status of gene B, or vice versa. Briefly, RNA expression of gene B was compared between cell lines where

gene A was essential versus nonessential using a one-sided Wilcoxon test and adjusted for multiple testing, with a threshold of $p<0.1$ for significance. In addition, the relationship between gene A essentiality and a loss-of-function alteration in gene B was tested using a one-sided Fisher test (false discovery rate (FDR) <0.1). Loss-of-function alterations in cell lines and cancer samples were defined as homozygous deletions or damaging mutations. These were annotated as described for cancer samples in Chapter 2.2.2. Of note, all 7,953 cancer samples were considered, with damaging alterations observed in 7,921 samples (Table 7-2).

2.4 Validation of synthetic lethal interactions

2.4.1 Cell culture

Cell lines were obtained from the Crick Cell Services Science Technology Platform, where they were authenticated and screened for mycoplasma (Table 7-3). Frozen cell vials were thawed in a 37°C water bath, diluted in 10ml medium and pelleted at 70g for 5min at 4°C. The pellet was resuspended in fresh medium and cells were grown at 37°C, 5% CO_2.

Cells were grown in T25 (Corning, Table 7-4) or T75 cell culture flasks (Thermo Fisher, Table 7-4) at 37°C, 5% CO_2 and media conditions as recommended for each cell line (Table 7-3). Cells were passaged upon reaching approximately 70-90% confluency. For passaging, cells were washed with phosphate buffered saline (PBS, Crick media preparation science technology platform, Table 7-4) once and trypsinised with 1ml/25cm^2 trypsin (Crick Media Preparation science technology platform, Table 7-4) at 37°C until cells detached. The trypsin was stopped with the same amount of medium and an aliquot of cells was transferred to a flask with fresh medium.

To prepare aliquots for freezing, cells were seeded into a T75 flask and grown to 70-90% confluency. Cells were trypsinised and the cell suspension was pelleted at 70g. The supernatant was removed, and the pellet was resuspended in 3ml freezing medium (90% foetal calf serum (FCS), 10% Dimethyl Sulfoxide (Sigma)).

This suspension was divided into 3 freezing vials for recovery into T25 flasks, and frozen at -80°C with a controlled cooling rate of 1°C/min. For long term storage, frozen vials were transferred to -150°C.

2.4.2 siRNA and CRISPR screen

The most promising synthetic lethal candidates among the top nine candidate pairs were identified with an siRNA knockdown and CRISPR Cas 9 knockout screen. This screen was performed in collaboration with the High Throughput Screening Science Technology Platform at the Crick.

Hacat and HEK293 cells were seeded to reach 70% confluency, trypsinised and counted. For the siRNA screen, library plates (siGENOME siRNA pool library, Horizon Discovery) were thawed and 10µl siRNA (equivalent to 3.75pmol) were transferred to a 96-well plate (Greiner Bio-One). As a transfection reagent, 0.2µl Lipofectamine™2000 (Thermo Fisher, for HEK293) or 0.2µl INTERFERin® (Polyplus, for Hacat) were diluted in 10µl Opti-MEM™ (Thermo Fisher) and added to the 96 well plate. Resulting in a final volume of 100µl per well, 6000 cells (HEK293) or 8000 cells (Hacat) per well were diluted in 80µl culture medium and added to the plate. Cells were incubated at 37°C and 5% CO_2 using an Incucyte® ZOOM system at 10X magnification (Sartorius) and confluency was measured over time using the corresponding Incucyte® ZOOM analysis software (Sartorius). Measurements were done in technical triplicates.

The CRIPSR Cas9 screen was performed in a similar way using a custom combined Edit-R™ crRNA library and Edit-R™ tracrRNA library (Horizon Discovery). Hacat and HEK293 cell lines with stable Cas9 expression were obtained from the Crick High Throughput Screening science technology platform. Data analysis including growth curve fitting and growth rate calculation was performed using the R grofit package (Kahm et al., 2010).

2.4.3 RNA extraction and qPCR

To determine expression levels of genes, RNA was extracted, reverse transcribed and a qPCR was performed. For RNA extraction, cells were trypsinised and an aliquot was pelleted at 900g and 4°C for 5min. The pellet was resuspended in 250µl lysis buffer containing 1% 2-Mercaptoethanol (Sigma) and RNA was extracted with the GenElute™ Mammalian Total RNA Miniprep kit (Sigma). RNA concentration was measured with a NanoDrop™ 1000 Spectrophotometer (Thermo Fisher) and a reverse transcription was performed using the High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher). In specific, 2µl 10x RT buffer, 2µl 10x RT random primers, 0.8µl dNTP mix, 1µl RT, 3.7µl ribonuclease (RNase) inhibitor (Thermo Fisher) and 3.7µl Nuclease-free Water (Omega Bio-tek Inc.) were mixed and added to 10µl RNA containing a total of 1000ng RNA. The mix was incubated in the MiniAmp™ Thermal Cycler (Thermo Fisher, pre-warm at 25°C for 10sec, incubate at 37°C for 2 hours, incubate at 85°C for 5min, hold at 4°C).

For each qPCR reaction, 5µl PowerUp™ SYBR™ Green Master Mix (Thermo Fisher), 0.4µl forward and reverse primer each (equivalent to 4pmol, Table 7-5) and 3.2µl Nuclease-free Water were mixed and pipetted on ice into a FrameStar® 384well PCR plate (4titude). cDNA was diluted 1:3 with nuclease free water and 1µl was added to the plate. Measurements were done in technical triplicates. The qPCR was run in a ViiA™7 thermal cycler (Thermo Fisher) with the following cycles:

- 2min at 50°C warm-up
- 2min at 95°C melting
- 40 cycles of: 1sec at 95°C melting, 20sec at 60°C annealing and elongation.

The QuantStudio™ Real-Time PCR Software (v1.2, Thermo Fisher) was used to record results. The ΔΔCT method was used to normalize results to the housekeeping gene (GAPDH).

2.4.4 CRISPR Cas9 editing of individual gene pairs

CRISPR Cas9 editing was achieved through introduction of gRNAs and recombinant Cas9 protein into the cell via electroporation. Cells were seeded to reach 70% confluency, trypsinised and counted. Per nucleofection, 500,000 cells were pelleted at 80g and 4°C for 5min, washed with 1ml PBS and pelleted again (80g, 4°C, 5min). Cells were resuspended in 10µl buffer R (Neon™ Transfection System 10µl Kit, Thermo Fisher). 9µl of the cell suspension were gently mixed with 3µl sgRNAs (equivalent to 90pmol, Synthego, Table 7-6) and 0.6µl TrueCut™ Cas9 Protein v2 (Invitrogen, equivalent to 3µg) and incubated for 10min at room temperature. Cells were electroporated using the Neon™ Transfection system (Thermo Fisher) and the Neon™ Transfection System 10µL Kit (Thermo Fisher) at cell line specific nucleofection conditions (Table 7-3). After nucleofection, cells were grown in 2ml pre-warmed medium in a 6-well plate (Corning, Table 7-4) at 37°C, 5% CO_2.

2.4.5 Validation of CRISPR Cas9 editing through Sanger sequencing

At several time points after nucleofection, the proportion of edited alleles was determined through Sanger sequencing. Briefly, cells were trypsinised and an aliquot of cells was pelleted (900g, 5min, 4°C). The supernatant was removed and genomic DNA was extracted using the PureLink™ Genomic DNA Mini Kit (Thermo Fisher). The concentration of DNA was measured using a NanoDrop™ 1000 Spectrophotometer (Thermo Fisher).

To amplify the edited region of interest, a polymerase chain reaction (PCR) was run using 25µl Q5® High Fidelity 2X Master Mix (New England BioLabs), 2.5µl forward and reverse primer each (equivalent to 25pmol, Table 7-5) and 20µl DNA (6ng/µl). The following PCR amplification cycles were applied using a MiniAmp™ Thermal Cycler (Thermo Fisher):

- 30sec at 98°C melting
- 35 cycles of: 10sec at 98°C melting, 15sec at 60°C annealing, 20sec at 72°C elongation
- 2min at 72°C final elongation
- 4°C hold

The purity of the PCR product was confirmed by gel electrophoresis on an agarose gel (1.5% UltraPure™ Agarose (Thermo Fisher) in tris acetate ethylenediaminetetraacetic acid (TAE) buffer (Crick media preparation science technology platform, Table 7-4) containing 1/5000 Nancy-520 (Merck)) and imaged using a UVP BioDoc-It® UV Transilluminator (Analytik Jena). The PCR product was further purified using the Monarch® PCR and DNA Cleanup Kit (5µg) (New England BioLabs).

Sanger sequencing was performed by the Crick Genomics equipment park science technology platform, Source Bioscience or using the OverNight Mix2Seq Kit (Eurofins Genomics) and sequences were analysed using the ICE v2 CRISPR analysis tool (Synthego).

2.4.6 Proliferation assay

To evaluate the effect of CRISPR Cas9 editing on cells, proliferation assays were performed. Briefly, cells were seeded to reach 70% confluency, trypsinised and counted. Cells were seeded into a 96well plate (Nunclon Delta Surface, Thermo Fisher, Table 7-4) at 100µl/well and at an optimized cell density (Table 7-3) and incubated at 37°C and 5% CO_2. At each time point, one plate was fixed with 50µl per well 4% formaldehyde (Thermo Fisher) in PBS for 10min, then formaldehyde solution was removed, 100µl PBS per well were added and the plate was stored at 4°C. At the last time point, all plates were stained with Crystal Violet solution (0.5g Crystal Violet (Bio Basic), 80ml Nuclease-free Water, 20ml methanol (Thermo Fisher)) at room temperature for 20min. The plate was then washed under tap water and dried overnight. To dissolve the Crystal Violet, 200µl methanol were added to each well and the plate was incubated at room temperature for 20min. The optical density at 570nm was measured using an

Infinite® F200 Pro plate reader (Tecan). Blank measurements were subtracted from each value, and the values were normalized to the beginning of the experiment. Measurements were done at least in triplicates.

2.4.7 Western Blot

A Western Blot was used to confirm effects of *MLLT1* and *MLLT3* loss on histone methylation, in specific H3K79 dimethylation (H3K79me2) and trimethylation (H3K79me3). For protein extraction, cells were seeded to reach 70% confluency, trypsinised and counted. One million cells were pelleted (100g, 4°C, 5min), frozen in liquid nitrogen and stored at -80°C. Protein was extracted from the pellet at room temperature using 100µl protein extraction buffer (33µl 3X Blue Loading Buffer (Cell Signalling Technology), 3.3µl dithiothreitol (DTT) (30X Reducing Agent (1.25M DTT), Cell Signalling Technology), 0.2µl Benzonase® Nuclease (Merck), 63.2µl nuclease-free water). Upon complete dissolution of the pellet, the solution was centrifuged to remove debris (16,200g, room temperature, 2min) and the supernatant incubated at 95°C for 5min. 20µl of the sample and 5µl of Precision Plus Protein™ Dual Color Standards (Bio-Rad) were loaded onto a TruPAGE™ Precast Gel 4-12% (Merck) and run at room temperature in TruPAGE™ SDS Running Buffer (Merck) at 125V for 1 hour. The proteins were then transferred to an X1 Amersham™ Protran™ 0.45µm Nitrocellulose membrane (GE Healthcare) in TruPAGE™ Transfer Buffer (Merck) + 20% methanol at 100V for 1 hour at 4°C. After transfer, the membrane was stained with Ponceau S staining solution (Torcis) to confirm successful transfer. The membrane was washed in tris buffered saline tween (TBST) washing buffer (1l tris buffered saline (TBS) (Crick Media preparation STP, Table 7-4) + 1ml Tween 20 (Thermo Fisher)) three times for 5min and blocked with 5% dried skim milk (Marvel) in TBST for one hour at room temperature. The membrane was incubated in the respective primary antibody dilution (Table 7-7) overnight. The membrane was washed three times for 5min in TBST and incubated for 1 hour at room temperature with the secondary antibody dilution (Table 7-7). The membrane was washed again three times for 5min and developed using the

Amersham™ ECL Western Blotting Analysis System (GE Healthcare) and an Amersham™ Imager 600 (GE Healthcare).

Chapter 3. The Network of Cancer Genes

3.1 Motivation

Finding effective treatments for cancer relies heavily on our understanding of its origin and progression. In this context, one of the main goals of cancer genomics has been the identification of genes whose somatic alterations play a role in tumour formation and progression, called cancer genes or cancer drivers. Advances in next generation sequencing of cancer genomes have increased our understanding of tumorigenesis. However, keeping an overview of cancer genes is becoming increasingly difficult. To address this issue, NCG collects a regularly updated list of cancer genes through manual curation of cancer sequencing screens (An *et al.*, 2016; An *et al.*, 2014; D'Antonio *et al.*, 2012; Dressler *et al.*, 2021; Repana *et al.*, 2019; Syed *et al.*, 2010). In addition, NCG annotates evolutionary properties of cancer genes, which are properties that distinguish them from the rest of human genes. Evolutionary properties annotated by NCG are gene duplicability, evolutionary origin, number of miRNA interactions, gene and protein expression in tissues, protein complex formation, and PPIN connectivity, centrality, and interconnectivity.

The following results focus on my contribution to two versions of the database, NCG6 (Repana *et al.*, 2019) and NCG7 (Dressler *et al.*, 2021). My contributions to NCG6 include the support of the literature curation and analysis of included screens (Figure 3-1). For both NCG6 and NCG7, I analysed PPIN characteristics, protein complex formation and gene essentiality (Figure 3-2A-C, part of Figure 3-3). For NCG7, I reviewed the literature regarding healthy driver genes and integrated a shiny application on the website to interactively display protein-protein interactions (Figure 3-4). The contributions of other group members to this work are also discussed and acknowledged throughout the results.

3.2 NCG6

3.2.1 Identification of canonical and candidate cancer genes

For NCG6 (Repana *et al.*, 2019), the identification of cancer genes from an extensive literature curation was led by my colleague Dimitra Repana, with contributions from other co-authors, including myself. The integration of two sources of known cancer genes, namely the Cancer Gene Census (Futreal *et al.*, 2004) and a publication by Vogelstein et al. (Vogelstein *et al.*, 2013), led to a total of 711 canonical cancer genes. This included 239 tumour suppressor genes, 239 oncogenes and 233 genes that could not be unambiguously classified either because the two sources did not agree in their annotation, or because of a proven dual role. In addition, we identified 1,661 candidate cancer genes through the curation of 273 cancer sequencing screens published between 2008 and 2018. Compared to the previous version, this amounted to a 1.5-fold increase in cancer genes, 98 additional publications, addition of seven new primary sites and 2.6-fold increase in donors (Figure 3-1A).

The large collection of cancer sequencing screens enabled us to identify which methods were predominantly used to distinguish alterations in driver genes from passengers. The most used method was the recurrent alteration of a gene within a patient cohort, without application of statistical methods to determine the recurrence threshold (Figure 3-1B). This approach is likely to lead to false positive identification of cancer drivers, as the threshold for recurrence is selected randomly by the respective study. Other frequently applied methods evaluated whether a gene was mutated more frequently in cancer samples than expected based on the background mutation rate. For example, MutSig (Lawrence *et al.*, 2013) corrects for variations in the background mutation rate by considering patient-specific mutation frequency and spectrum as well as gene-specific background mutation rates. Similarly, MuSiC (Dees *et al.*, 2012) includes options to calculate the background mutation rate for a whole group of samples or a sample-specific background. All these methods rely on a large sample cohort to identify cancer genes, especially if they are rarely altered. Therefore, it was not surprising that we found a positive correlation between the number of cancer

donors in a study, and the number of cancer genes identified (Figure 3-1C). Finally, recent studies have started to use multiple prediction methods (Figure 3-1D), potentially leading to more robust cancer gene identification in the future.

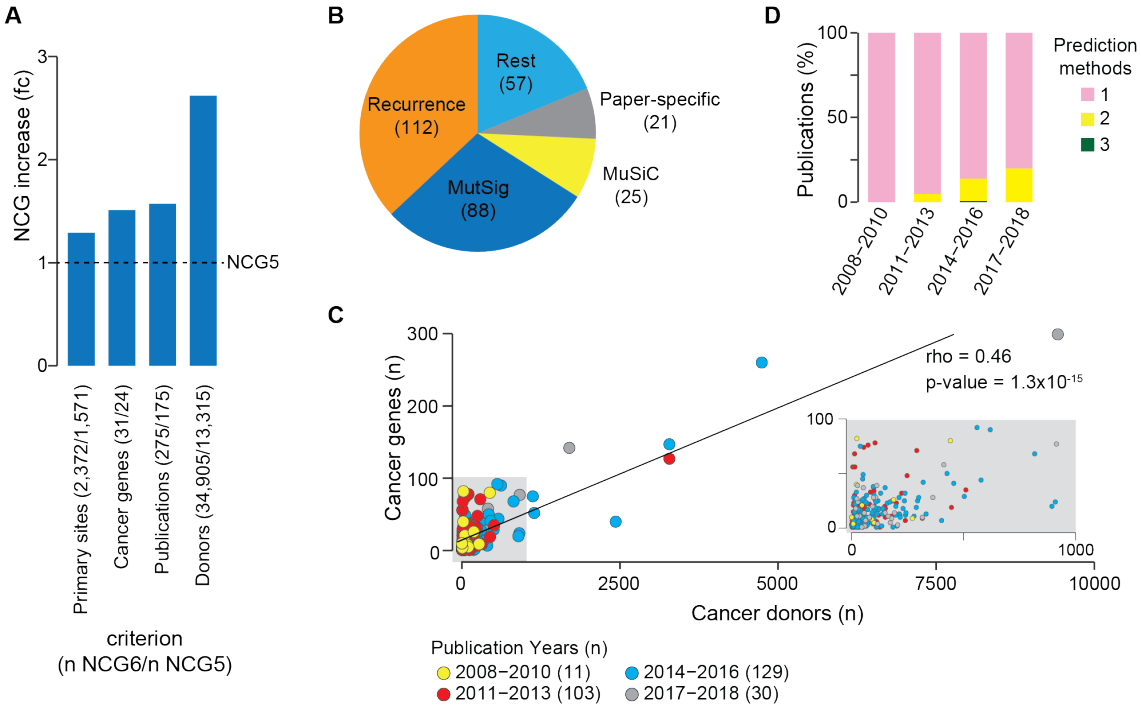


Figure 3-1 Literature curation of NCG6.

A) Comparison of included primary sites, cancer genes, publications and donors between NCG5 and NCG6. fc: fold change **B)** Methods applied for the identification of cancer genes in 273 publications. Numbers of studies for each method are indicated in brackets. Some studies used more than one method and were included in several categories. **C)** Cancer donors and cancer genes identified per study. The grey insert zooms into the bottom left corner of the plot. The correlation was calculated using the Spearman correlation. **D)** Number of methods used per publication to identify cancer genes. One study applied the method PanSoftware, which was considered as one method but actually combines 26 prediction tools.

This figure was adapted from Repana et al. (Repana *et al.*, 2019).

3.2.2 Evolutionary properties of canonical and candidate cancer genes

The evolutionary history of cancer genes has given them distinct properties compared to the rest of human genes (An *et al.*, 2016; D'Antonio and Ciccarelli, 2011; Domazet-Loso and Tautz, 2010; Jonsson and Bates, 2006; Xia et al., 2011). To identify these properties, we first confirmed previously described properties, and then expanded them using large scale datasets that had recently become available. Based on these datasets we only included new properties if information was available on the majority of human genes and if we observed a significantly different signal for cancer genes compared to the rest of human genes. We compared 711 canonical cancer genes and 1,661 candidate cancer genes from the literature with the rest of human genes. We also compared 239 tumour suppressors with 239 oncogenes and investigated a subset of 104 high support candidate cancer genes that were validated in at least two independent sequencing screens. We hypothesized that these high support candidates were more likely to be true cancer genes.

We confirmed that proteins encoded by canonical cancer genes were more connected, central, and inter-connected in the human PPIN compared to the rest of human proteins (Figure 3-2A). This trend was more pronounced for high support candidates than for all candidate cancer genes. We confirmed that proteins encoded by canonical and candidate cancer genes, especially those with high support, were more frequently involved in protein complexes (Figure 3-2B). Furthermore, we expanded the catalogue of cancer gene evolutionary properties by adding gene essentiality in cancer cell lines as identified by the OGEE (Chen *et al.*, 2017) and PICKLES (Lenoir *et al.*, 2018) databases. Canonical cancer genes and high support candidates were more often essential in at least one cell line compared to the rest of human genes (Figure 3-2C). Canonical cancer genes were also essential in more cell lines than the rest of human genes, while tumour suppressors were essential in more cell lines than oncogenes (Figure 3-2C). We found that a significantly lower fraction of tumour suppressors had duplicated copies in the human genome compared to oncogenes but did not identify a significant difference between cancer genes overall and the rest of human genes (Figure 3-2D). Candidate cancer genes originated earlier in evolution than the

rest of human genes, while canonical cancer genes were composed of evolutionary older tumour suppressors and younger oncogenes (Figure 3-2E). Regarding expression in healthy human tissues, canonical cancer genes were more broadly expressed on RNA (Figure 3-2F) and protein (Figure 3-2G) level, with tumour suppressors expressed more broadly than oncogenes. Canonical cancer genes were targeted by more miRNAs than the rest of human genes and were more central in the miRNA interaction network. This observation was also true for candidate cancer genes, with a greater difference to the rest of human genes for candidates with high support (Figure 3-2H). In addition, cancer genes were enriched in certain functional pathways such as signal transduction, chromatin organization and cell cycle, and depleted in others such as metabolism and transport (Figure 3-2I,J). Candidates generally exhibited a weaker enrichment and depletion, except for extracellular matrix organization (Figure 3-2I) where they were enriched but canonical cancer genes were not. Interestingly, canonical cancer genes were enriched in the transcription pathway in Reactome, but no cancer gene group was significantly enriched in this pathway using KEGG annotations.

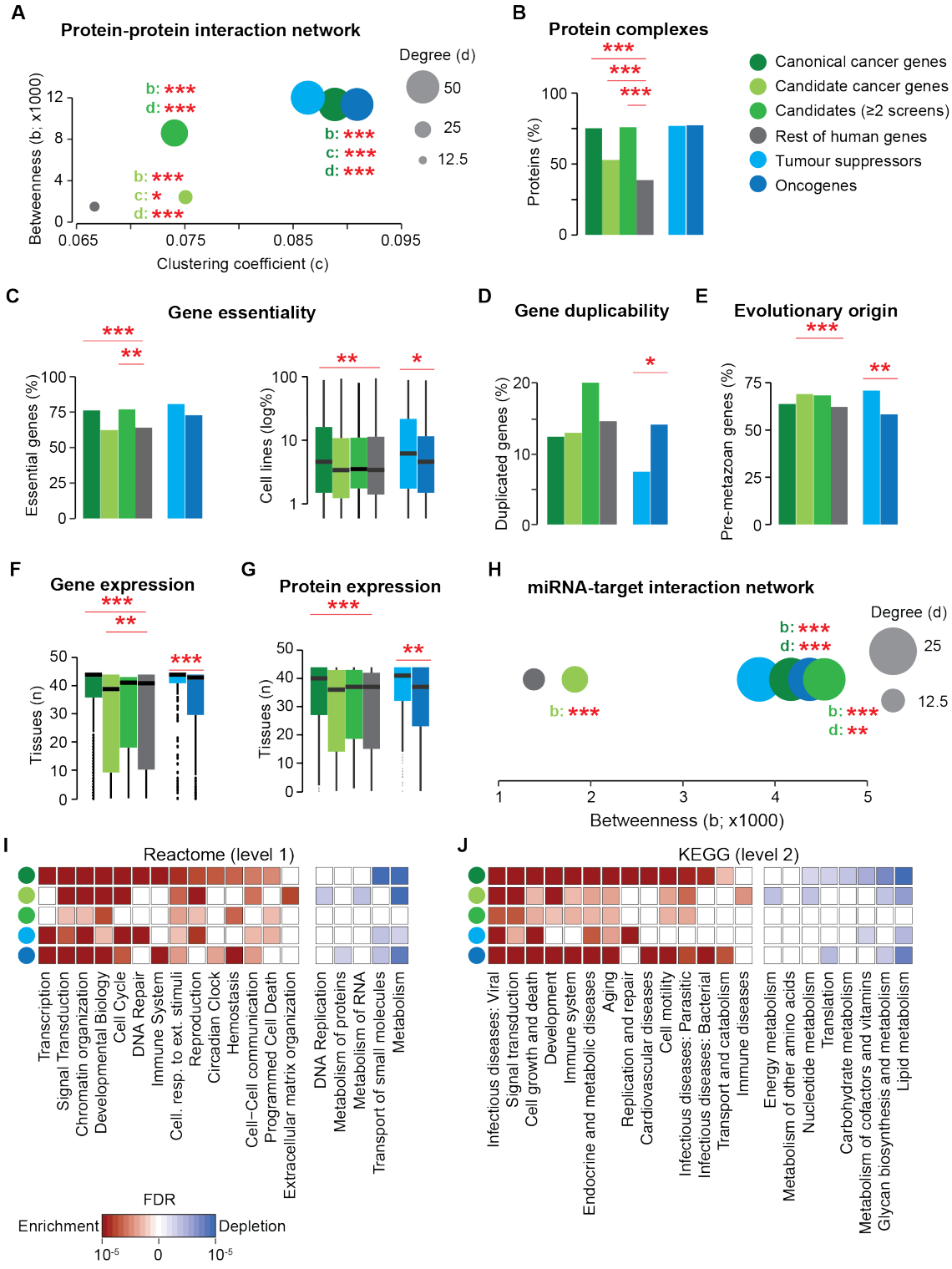


Figure 3-2 Evolutionary properties of cancer genes.

A) Median values of betweenness (centrality), clustering coefficient (inter-connectivity) and degree (connectivity) of protein groups in the human PPIN. **B)** Percentage of proteins involved in at least one protein complex. **C)** Percentage of genes which are essential in at least one cell line, and distribution of number

of cell lines in which genes are essential. Genes were only included if their annotation agreed in OGEE and PICKLES. **D)** Percentage of genes with at least one duplicate (>60% coverage of the protein sequence) in the human genome. **E)** Proportion of genes with pre-metazoan origin. **F)** Number of human tissues in which RNA is expressed. Only genes with expression annotation in both GTEx and Protein Atlas were included, and tissue types were matched between both sources if possible. Genes were defined as expressed if they had >1TPM expression in both datasets. **G)** Number of human tissues in which protein is expressed. **H)** Median values of betweenness and degree of gene groups in the human miRNA-target interaction network. Since interactions only occur between miRNAs and target genes, the clustering coefficient is always 0. In all panels, canonical and candidate cancer genes were compared to the rest of human genes, and tumour suppressors were compared to oncogenes. Significance was calculated using a two-sided Fisher test (B, C, D, E) or Wilcoxon test (A, C, F, G, H). *p<0.05, **p<0.01, ***p<0.001. **I,J)** Enrichment and depletion of functional groups in cancer driver categories (I: Reactome level 1, J: KEGG level 2). Significant differences were calculated by comparing the respective cancer gene groups to the rest of human genes using a two-sided Fisher test and calculating false discovery rates for each gene separately. Only pathways with enrichment or depletion are shown.
This figure was adapted from Repana et al. (Repana *et al.*, 2019).

Overall, these results confirmed that cancer genes have distinct properties. Within canonical cancer genes, differences were also present between tumour suppressors and oncogenes. Candidate cancer genes followed the same trends as canonical cancer genes. Interestingly, candidates supported by at least two independent cancer sequencing screens were more similar to canonical cancer genes compared to the rest of candidates.

3.3 NCG7

3.3.1 Additional driver gene categories in NCG7

In the seventh release of NCG (Dressler *et al.*, 2021), we added a third resource of canonical cancer drivers (Saito *et al.*, 2020), 32 screens investigating alterations in coding regions of the genome and 19 screens focussing on alterations in noncoding regions of protein-coding genes. In addition, my colleague Neshika Wijewardhane and I curated 18 publications regarding the identification of genes whose alterations drive clonal expansion in non-cancer tissue, called healthy drivers. Overall, this led to 591 canonical cancer drivers including 256 oncogenes and 254 tumour suppressors and 2,756 candidate cancer drivers. The decrease of canonical drivers compared to NCG6 resulted from the more stringent inclusion criteria from the Cancer Gene Census (Futreal *et al.*, 2004). Most of the 3,177 drivers identified through sequencing screens were altered in coding regions, only 531 were altered in noncoding regions of protein-coding genes and 190 in both. We identified a high overlap of cancer and healthy drivers. Out of 95 identified healthy drivers, 57 were also annotated as canonical cancer drivers and 30 as candidate cancer drivers.

3.3.2 Evolutionary properties of cancer and healthy drivers

Similar to NCG6, we analysed evolutionary properties of cancer genes, and added the prevalence of germline variation to the property repertoire. We confirmed characteristic evolutionary properties of cancer genes, with canonical cancer drivers having stronger differences from the rest of human genes than candidates (Figure 3-3A). We also confirmed differences between tumour suppressors and oncogenes (Figure 3-3B) and observed that the main difference of candidates was driven by coding candidates, while non-coding candidates closely resembled the rest of human genes. (Figure 3-3C).

Genes that were both canonical cancer and healthy drivers had the most pronounced differences from the rest of human genes, while candidate cancer and healthy drivers had a weaker, yet similar, property profile (Figure 3-3D).

Healthy drivers that were never observed as cancer drivers did not significantly differ from the rest of human genes, except for their lower tolerance to germline variation (Figure 3-3D). However, this group was only composed of eight genes, and identification of further remaining healthy drivers may refine their characterisation.

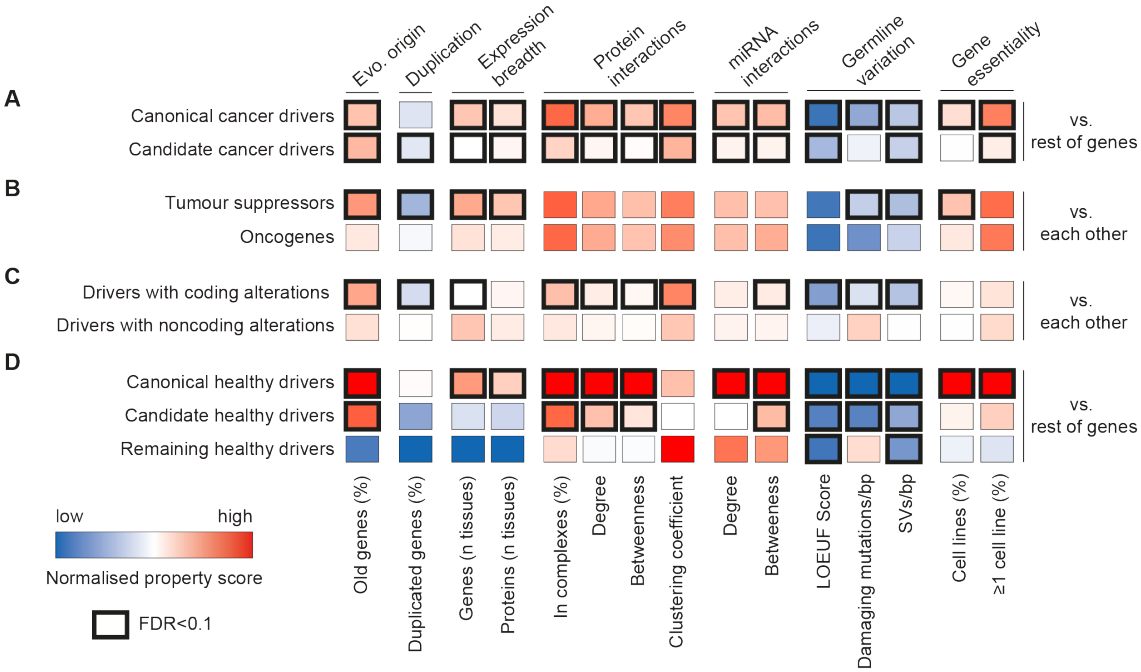


Figure 3-3 Evolutionary properties of driver gene groups.

Comparison of evolutionary properties between **A)** canonical or candidate cancer drivers versus the rest of human genes; **B)** tumour suppressors versus oncogenes; **C)** drivers with coding alterations versus drivers with noncoding alterations; **D)** canonical, candidate and remaining healthy drivers versus the rest of human genes.

This figure was adapted from Dressler et al. (Dressler *et al.*, 2021). The original figure was produced by my colleague Giulia Sartini as a summary of results contributed by co-authors of the study including myself.

3.3.3 Interactive display of protein-protein interactions on the website

As part of the update to NCG7, we improved the NCG website (www.network-cancer-genes.org). This included the integration of a shiny application to interactively display protein-protein interactions. The application contains

interaction data from the NCG database and displays interactions between proteins with properties defined by the user (Figure 3-4).

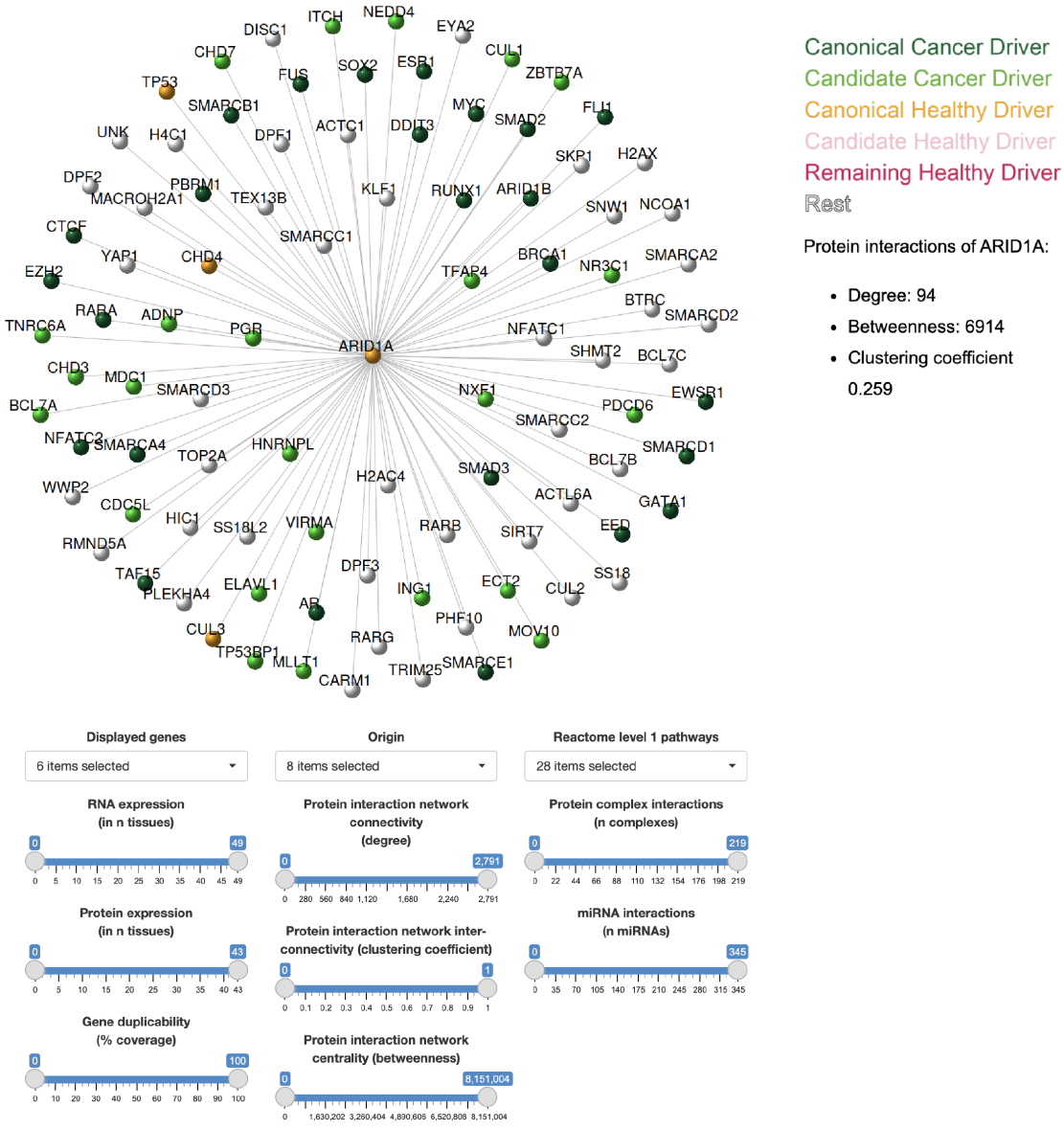


Figure 3-4 Interactive PPIN display on the NCG7 website.

The selected gene is displayed in the middle, with interactors conforming to selection criteria in the periphery. Genes are colour coded according to their driver gene status. Users can select different criteria to filter the displayed interaction partners.

3.4 Conclusion

The Network of Cancer Genes provides a comprehensive resource of canonical and candidate cancer genes. It has now been extended to cancer genes altered in non-coding regions and drivers of clonal expansion in non-cancer tissues. I was involved in the publication of two releases, NCG6 and NCG7. In addition to confirming known evolutionary properties of cancer genes, I introduced the essentiality of cancer genes as an additional property, contributed to the inclusion of healthy drivers and expanded the NCG website. This extensive collection of genes and their properties represents an informative resource for cancer researchers, which is available online at www.network-cancer-genes.org.

The collection of cancer sequencing screen publications over more than ten years enabled us to identify potential biases in the identification of cancer genes. First, the most frequently used method to identify cancer genes was the recurrence of certain alterations, where recurrence was defined by a publication-specific threshold. This may have biased our collection in favour of genes discovered by studies applying a less stringent threshold. However, it is reassuring to see an increasing number of studies using several methods to verify identified cancer genes. Even studies using sophisticated models to identify thresholds of recurrence are limited by the number of samples available. The positive correlation between the number of donors and the number of identified cancer genes per study indicates that further large-scale cancer sequencing screens may be able to identify drivers with lower recurrence frequency.

Cancer genes are characterized by distinct evolutionary properties. Compared to the rest of human genes, these properties are most pronounced in those genes which drive both cancer and healthy clonal expansion. In contrast, they are less pronounced in candidate cancer genes, possibly because candidates are generally more similar to the rest of human genes, or because the group of candidates is a mix of true positive and false positive cancer genes. Supporting the latter hypothesis, the group of candidates with high support from the literature resembles canonical cancer genes.

Chapter 4. Evolutionary properties of biological pathways in health and disease

4.1 Motivation

Evolutionary properties of genes are connected to gene function in health and disease. As described in Chapter 3, canonical cancer genes differ from the rest of human genes regarding their evolutionary properties. Candidate cancer genes resemble canonical cancer genes, but with less pronounced differences to the rest of human genes, possibly due to the inclusion of false positives. In addition, some distinctive Mendelian disease gene properties have been described, for example early origin in evolution (Domazet-Loso and Tautz, 2010) or their position in the PPIN (Dickerson *et al.*, 2011; Goh *et al.*, 2007). Independent of their role in diseases, genes involved in the same biological process are also characterized by similar evolutionary properties (Castellana *et al.*, 2018; Li and Zhang, 2017; Sauka-Spengler *et al.*, 2007; Szedlak *et al.*, 2016).

Evolutionary properties are also connected to each other (Chen *et al.*, 2020; Chen *et al.*, 2012; D'Antonio and Ciccarelli, 2011; Freilich *et al.*, 2005; Hart *et al.*, 2014; Hughes and Friedman, 2005; Liang and Li, 2007a; Liao and Zhang, 2007; Makino *et al.*, 2009; Papp *et al.*, 2003; Prachumwat and Li, 2006; Rambaldi *et al.*, 2008; Veitia, 2002; 2004; Wang *et al.*, 2015; Yang *et al.*, 2003). Since a comprehensive analysis of properties does not exist, the collective interplay between evolutionary properties and gene function in health and disease remains unknown. The following questions are therefore addressed in this chapter:

1) How is the gene function connected to a gene's evolutionary path, in specific to gene evolutionary origin, breadth of mRNA and protein expression across tissues, duplicability, number of miRNA interactions, number of protein complexes and PPIN connectivity, centrality and inter-connectivity?
2) How are gene evolutionary properties connected to a gene's tolerance to germline and somatic loss of function as well as diseases caused by genomic alterations?

3) Are false positives the reason why candidate cancer genes exhibit weaker differences from the rest of human genes than canonical cancer genes?

4.2 Inter- and intra-pathway heterogeneity of gene evolutionary properties

To identify the connection between gene function and evolutionary properties, we characterized nine evolutionary properties of 19,549 human genes and proteins (Methods, Table 4-1) across functional pathways. Given the distinct evolutionary properties of cancer genes (Chapter 3.2), we reasoned that these properties were likely to differentiate functional pathways as well. We removed gene essentiality from the list of properties to avoid a bias when investigating properties of essential genes (Chapter 4.4).

Of 19,549 investigated genes, 10,334 were members of 25 functional pathways as annotated by Reactome level 1 pathways (Fabregat *et al.*, 2018).

Table 4-1 Gene evolutionary properties.

For each property, the number of genes is indicated for which data were available and for which Reactome Level 1 pathways (Fabregat *et al.*, 2018) were annotated.

EP	Description	Unit compared in Figure 4-1	Genes (n)	Genes in Reactome (n)
Gene evolutionary origin	Oldest ancestor with an orthologous gene	Pre-metazoan genes (%)	18,486	10,197
Gene duplicability	Retention of gene duplicates sharing >60% of the protein sequence	Duplicated genes (%)	19,549	10,334
Breadth of mRNA expression	mRNA expression in 43 healthy tissues	Tissues (n)	18,641	10,244
Breadth of protein expression	Protein expression in 44 healthy tissues	Tissues (n)	13,001	7,596
miRNA-gene interactions	Number of interacting miRNAs	Interactions (n)	14,649	8,411
Engagement in protein complexes	Number of protein complexes the protein participates in	Complexes (n)	8,080	7,231
PPIN connectivity	Number of direct neighbours	Degree (n)	16,322	9,377
PPIN centrality	Involvement in n shortest paths between two proteins	Betweenness (n)		
PPIN inter-connectivity	Fraction of existing/ all possible connections between direct neighbours	Clustering coefficient (n)		
Total	**Unique gene loci**	**NA**	**19,549**	**10,334**

Except for PPIN inter-connectivity, the properties of genes included in Reactome differed significantly from those of genes not annotated in Reactome (Figure 4-1). This may indicate a bias in property annotation for well-known genes and proteins.

For example, they may be preferentially included in screens investigating protein complex interactions, miRNA interactions or targeted protein interaction screens. We ordered pathways by their proportion of genes with pre-metazoan origin to estimate the extent to which evolutionary origin had an impact on evolutionary properties. While we observed parallels between evolutionary age and some properties, for example breadth of expression, the connection to other properties such as participation in complexes was not as obvious.

For all nine properties, we observed a considerable heterogeneity between pathways (Figure 4-1). For example, the pathways DNA replication or metabolism of RNA included genes which tend to originate early in evolution (Figure 4-1A) and had duplicated loci in the genome (Figure 4-1B). They were also broadly expressed across tissues at mRNA (Figure 4-1C) and protein (Figure 4-1D) levels, were targets of several miRNAs (Figure 4-1E) and encoded proteins involved in multiple protein complexes (Figure 4-1F). Finally, they were highly connected, central, and inter-connected in the PPIN (Figure 4-1G-I). Genes associated with digestion and absorption, or functions of the neuronal system had an opposing property profile. They were preferentially young, had a low tendency to duplicate, showed a tissue specific expression, were targeted by few miRNAs and encoded peripheral proteins in the PPIN that engaged in few complexes.

In addition to inter-pathway variability, we observed intra-pathway heterogeneity. For example, reproduction-related proteins varied in the number of protein complexes they participated in (Figure 4-1F) and genes involved in signal transduction, cell-cell communication and development were expressed in a variable number of tissues (Figure 4-1C,D). This indicated that different gene groups had distinct evolutionary properties which were dependent on each other to a certain extent as described in the literature (Chapter 1.2). However, due to the high number of properties and pathways, as well as intra-pathway heterogeneity, obtaining a comprehensive overview of pathway evolutionary properties and identifying similarities between pathways was challenging.

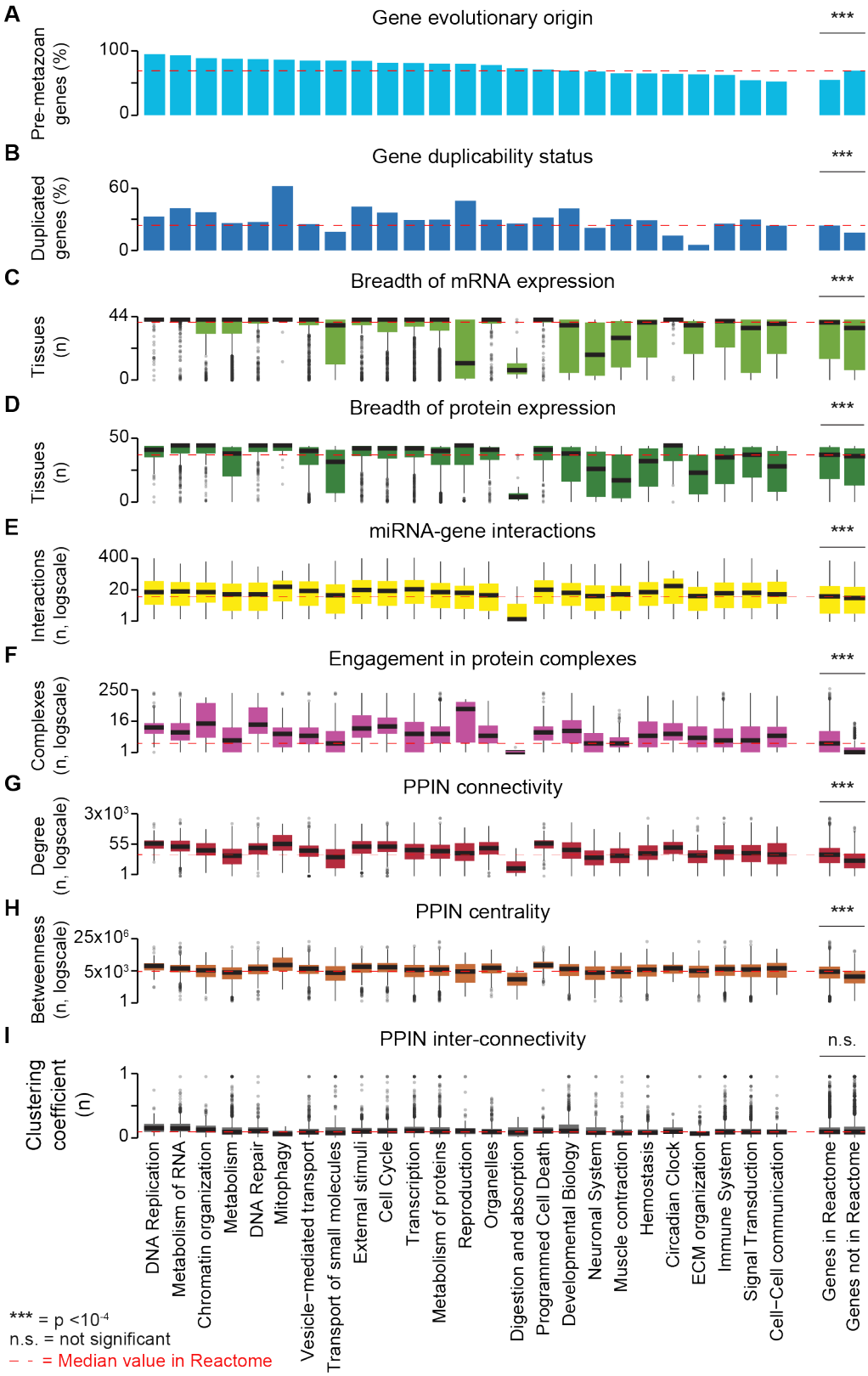


Figure 4-1 Evolutionary properties of human functional pathways.

For each of the 25 human functional pathways, the following evolutionary properties are displayed: **A)** the percentage of genes with a pre-metazoan origin,

B) the percentage of genes duplicated in the human genome, the breadth of **C)** mRNA and **D)** protein expression in human tissues, **E)** the distribution of miRNAs interacting with pathway-associated genes, **F)** the distribution of complexes that pathway-associated proteins participate in, the distribution of the PPIN **G)** connectivity (degree), **H)** centrality (betweenness) and **I)** inter-connectivity (clustering coefficient) of pathway-associated proteins. p-values were calculated using a two-sided Fisher test (A, B) or a two-sided Wilcoxon rank sum test (C-I). Shortened pathway names: External stimuli - Cellular responses to external stimuli, Organelles - Organelle biogenesis and maintenance, Transcription - Gene expression (Transcription), ECM organization - Extracellular matrix organization.

4.3 Gene evolutionary properties divide functional pathways into three groups

To dissect the interplay between evolutionary properties and gene function more broadly, we integrated the nine properties into one score. This was done by training a random forest classifier using 148 and 117 genes with the top and bottom 25% values of all numerical properties as a positive and negative training set, respectively (Figure 4-2A). The trained model estimated the likelihood of a gene to belong to the positive training set, which we called the EP score. Thus, a gene with consistently high numerical property values was similar to the positive training set, and therefore had a high EP score. Except for the number of complexes that the protein engages in, all numerical evolutionary properties contributed substantially to the EP score, whereas the two categorical properties, evolutionary age and duplicability, had little impact (Figure 4-2B). This was expected, as they were not considered for the creation of the training sets.

Comparing the EP score distributions across functional pathways revealed three distinct groups (Figure 4-2C). Group 1 had a median EP score >0.8 and was composed of pathways related to basic cell functions, such as DNA repair and replication, RNA metabolism, cell cycle and programmed cell death. Group 2 showed a median EP score between 0.6 and 0.8 and included diverse pathways

such as transcription, metabolism of proteins, immune system and development. Finally, group 3 had EP scores <0.6 and included pathways important for multicellular organisms and organ specific functions, such as extracellular matrix organisation and muscle contraction. To avoid a biased signal based on few genes participating in several pathways, we confirmed these results after excluding genes participating in more than two pathways (Figure 4-2D).

We repeated the analysis using the independent pathway database KEGG (Kanehisa *et al.*, 2017), including 6,538 genes and five level 1 pathways. Similar to the Reactome group 1 pathways, we observed a high median EP score for genes involved in genetic information processing, a pathway representing basic cell functions (Figure 4-2E). Similarly, environmental information processing, metabolism and organismal systems resembled the Reactome group 3 pathways, and their median EP scores were below 0.6.

We ranked the nine properties by median value (numerical properties) or percentage (categorical properties) across 25 pathways to identify their contribution to the EP score. For group 1, the majority of properties ranked in the top ten positions. In contrast, properties of group 3 ranked in the bottom ten positions and group 2 had intermediate property ranks (Figure 4-2F). Overall, this indicated a good representation of most properties by the EP score. One exception was the reproduction pathway which had broadly distributed property ranks. Unsurprisingly, gene evolutionary origin and gene duplicability were the most frequent outliers (Figure 4-2F), as they were not used to determine the positive and negative training set for EP score calculation. Outliers were also present to a lesser extent regarding miRNA-gene interactions.

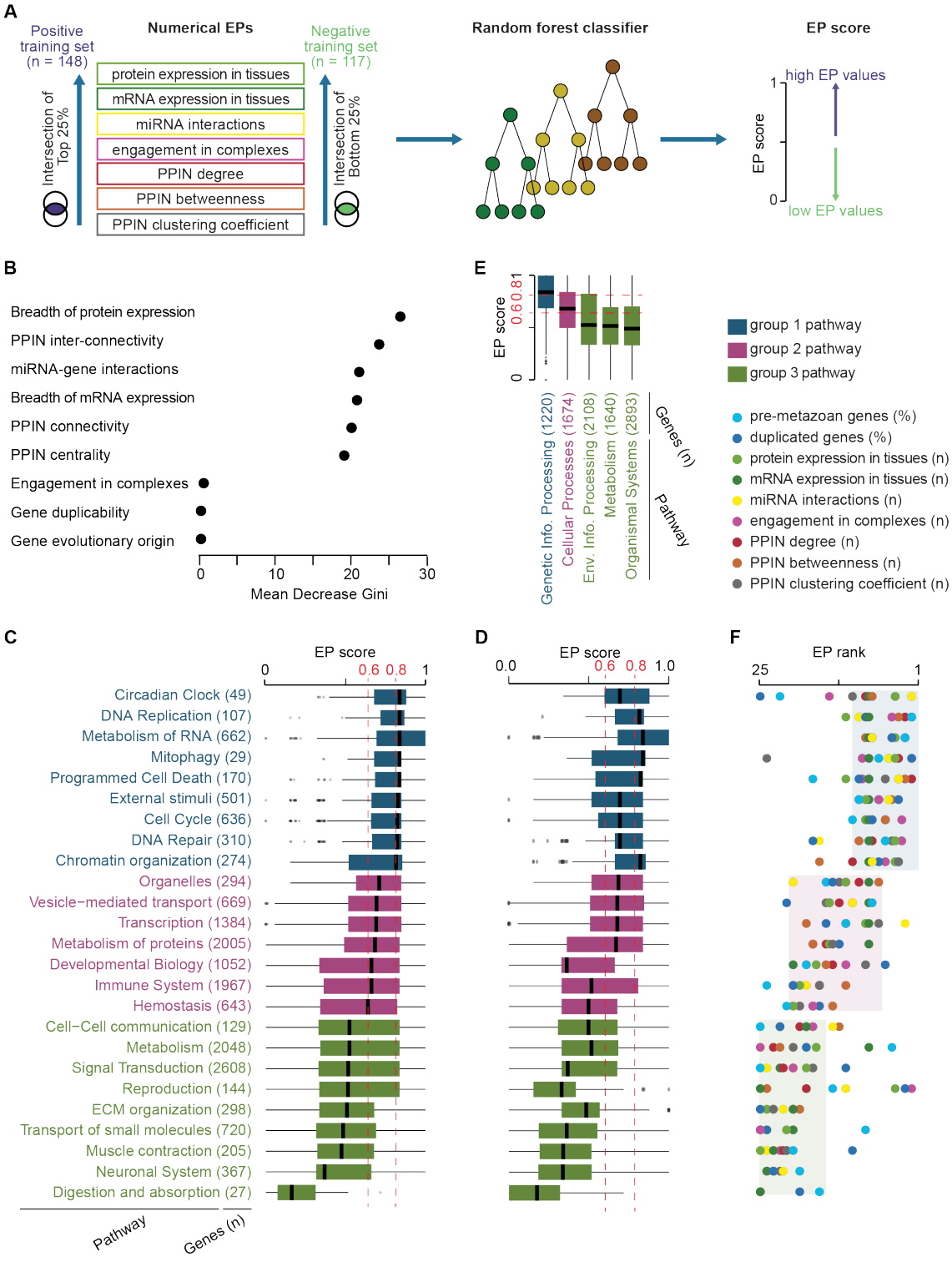


Figure 4-2 EP score distribution across pathways.

A) Workflow to derive the EP scores of 19,549 human genes. Genes with the top and bottom 25% of values of numerical evolutionary properties were used as a positive and a negative set to train a random forest classifier. The trained classifier was then used to measure the likelihood of a gene to be part of the

positive set (EP score). **B)** Contribution of nine properties to the EP score. The contribution was defined as the mean decrease in Gini value. It quantifies the extent to which each evolutionary property contributes to the homogeneity within the nodes and leaves of the random forest model. A high Gini decrease corresponds to a higher increase in homogeneity and thus higher impact of a property to the final Random Forest estimate and therefore the final EP score. **C-D)** Distribution of the EP score of C) all genes and D) pathway-specific genes associated with 25 Reactome level 1 pathways. In D), genes which were part of more than two pathways were removed from the analysis. Pathways were divided into three groups according to the median EP score (blue: pathway median > 0.8; pink: 0.6 < pathway median < 0.8, green: pathway median < 0.6). Shortened pathway names: External stimuli - Cellular responses to external stimuli, Organelles - Organelle biogenesis and maintenance, Transcription - Gene expression (Transcription), ECM organization - Extracellular matrix organization. **E)** EP score distribution of genes associated with five KEGG level 1 pathways. Pathways were divided into three groups as in C and D. Shortened pathway names: Env. – Environmental, Info. – Information. **F)** Ranking of nine evolutionary properties across the 25 pathways. Properties were ranked according to their median (numerical properties) or percentage (categorical properties) evolutionary property values in 25 pathways. Blue, pink and green boxes represent the typical rank range for pathways of the respective group.

While the pathway median EP score allowed a broad grouping of pathways based on their evolutionary properties, it limited the amount of information on the interplay between pathway function and evolutionary properties due to intra-pathway heterogeneity. To further dissect the contribution of each of the nine properties at a single gene level, we used a PCA. We excluded 43 outlier hubs with PPIN degree >500, resulting in 10,291 genes included in Reactome used for this analysis. Missing values for properties were imputed where needed (Methods). Together, Principal Components 1 and 2 explained approximately 45% of the variation (Figure 4-3A) of the resulting PCA distribution (Figure 4-3B). Gene evolutionary origin, breadth of mRNA and protein expression and number

of miRNA-gene interactions separated genes in the PCA plot in the same direction (Figure 4-3C,D). This suggested that genes with an early evolutionary origin tend to be broadly expressed and interact with several miRNAs. PPIN connectivity, centrality and engagement in protein complexes separated genes in an approximately orthogonal direction (Figure 4-3C,D). The majority of genes encoding protein hubs and subunits of many complexes were also old, broadly expressed and interacting with several miRNAs. We observed high EP scores for genes with high values of principal component 1, demonstrating concordance between the two methods (Figure 4-3D,E).

To investigate individual pathways, we projected the EP scores of their genes into the PCA. Genes in the same pathway tended to have similar EP scores and proximal locations in the PCA plot (Figure 7-1). Thus, despite the intra-pathway heterogeneity (Figure 4-1), evolutionary properties tended to be more similar within than across pathways (Figure 4-3F-H). In line with its broad distribution of evolutionary property ranks (Figure 4-2), the reproduction pathway was the only pathway that differed from this trend. It was formed of two groups of genes with distinct EP scores and locations in the PCA (Figure 7-1C). We reasoned that this was likely due to their specific function regarding reproduction: genes with a high EP score were often also involved in cell cycle and meiosis, an aspect of reproduction conserved across all organisms. In contrast, genes with a low EP score were related to spermatogenesis, similar to other low-scoring pathways that developed recently in multicellular organisms. In conclusion, the inter-and intra-pathway heterogeneity of evolutionary properties reflected specific biological functions and their origin in evolution.

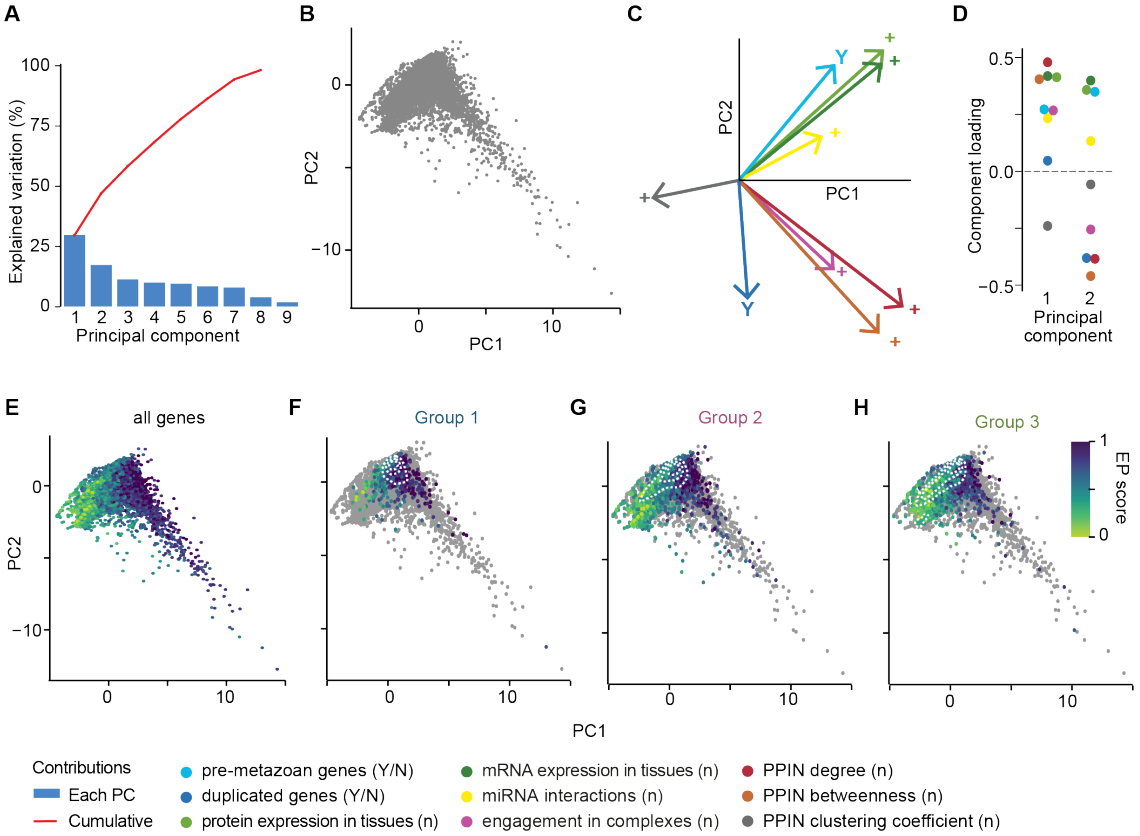


Figure 4-3 Principal component analysis of gene evolutionary properties.

A) Explained variation of the data per Principal Component. The PCA was performed on 10,291 Reactome genes and nine evolutionary properties. **B)** PCA including nine properties and 10,291 Reactome genes. The score plot shows contributions of the first and second principal components, each dot represents one gene. Forty-three outlier genes with PPIN degree >500 were removed. **C)** Contribution of each property to the first two principal components. The relative magnitude and direction of each property from the coordinate origin is shown. Arrows point towards areas where genes with higher numerical property values (+), duplicated genes (Y) and genes with pre-metazoan origin (Y) are located. **D)** Component loading for principal components 1 and 2. The component loading indicates the direction and magnitude of the contribution of each property to the Principal component. **E-H)** PCA including nine properties and 10,291 Reactome genes with EP score representation. The score plot was derived as in panel B, with dots coloured according to the gene EP score and showing E) all genes, genes belonging exclusively to F) pathway group 1, G) pathway group 2 or H)

pathway group 3. Genes belonging to multiple pathway groups are not included in any of the panels. Gene density lines are shown in white.

4.4 Pathway EP scores correlate negatively with tolerance to germline and somatic gene loss

Based on our interest in the relationship between evolutionary properties of genes and diseases caused by genetic alterations, we expanded our analysis to include tolerance towards germline and somatic gene loss-of-function alterations. We used the LOEUF score as a measure of tolerance to damaging germline alterations (Karczewski *et al.*, 2020). This score was obtained in a study of 125,748 healthy individuals and is calculated as the observed divided by the expected number of loss-of-function alterations per gene. Thus, it increases with increasing tolerance towards damaging alterations. We observed a negative correlation between median pathway EP and LOEUF scores, indicating a lower tolerance towards germline loss-of-function alterations in pathways with high EP scores (Figure 4-4A). Despite this, there was a high variability across pathways (Table 7-1). For example, the pathways DNA repair and mitophagy had a high EP and LOEUF score, meaning a high tolerance for damaging germline alterations (Figure 4-4A). The low-scoring pathways cell-cell communication and neuronal system had a low LOEUF score (Figure 4-4A).

We measured somatic alteration tolerance by the median number of genes per pathway acquiring loss-of-function alterations in 7,626 TCGA cancer samples (Table 7-2) or 1,234 cancer cell lines (Methods). Similar to germline alterations, the tolerance towards somatic alterations in cancer samples (Figure 4-4B) and cell lines (Figure 4-4C) decreased with increasing median pathway EP score, but we identified different outliers. For example, the high scoring pathways chromatin organization and circadian clock showed a high tolerance towards somatic loss of function. In addition, reproduction was the most prominent outlier among the low scoring pathways, possibly reflecting its heterogeneous composition.

Despite the known divergence between cancer samples and cancer cell lines, we observed a strong positive correlation between the number of loss-of-function

alterations in cancer cell lines and cancer samples (Figure 4-4D). While the median pathway EP score anti-correlated with both germline and somatic alteration tolerance, we did not observe a correlation between them (Figure 4-4E). This suggested a different selective pressure acting on germline and somatic mutations.

We investigated the relationship between median pathway EP scores and essentiality of the pathway-associated genes. To obtain essentiality information, we integrated datasets from six genome-wide CRISPR Cas9 knockout screens (Behan *et al.*, 2019; Dempster *et al.*, 2019; DepMap Broad, 2019a; b; 2020; Lenoir *et al.*, 2018; Meyers *et al.*, 2017) and three RNAi knockdown screens (Lenoir *et al.*, 2018; McFarland *et al.*, 2018; Tsherniak *et al.*, 2017) (Methods). We defined 444 genes as core essential (essential in at least 80% of tested cell lines), 4,480 genes as context dependent essential (essential in at least two cell lines and less than 80% of tested cell lines) and the remaining 13,587 genes as nonessential (Methods, Table 4-2).

Table 4-2 List of essential genes, cancer genes, healthy drivers and Mendelian disease genes.

Datasets on gene essentiality, cancer genes, healthy driver genes and Mendelian disease genes were downloaded and integrated. For each gene group, the number of genes for which data were available and for which Reactome Level 1 pathways were annotated is indicated.

Annotated functional group		**Source**	**Genes (n)**	**Genes in Reactome (n)**
Gene essentiality	All	Union of genes in DepMap (Behan et al., 2019; Dempster et al., 2019; DepMap Broad, 2019a, b, 2020; McFarland et al., 2018; Meyers et al., 2017; Tsherniak et al., 2017) and PICKLES (Lenoir et al., 2018)	18,511	10,241
	core essential (essential in >80% of tested cell lines)		444	375
	context essential (essential in >1 cell line and <80% of tested cell lines)		4,480	2,906
	nonessential (essential in 0 or 1 tested cell line)		13,587	6,960
Cancer genes	All canonical cancer drivers	Network of cancer genes (Repana et al., 2019)	711	533
	Tumour suppressors		239	191
	Oncogenes		239	181
	Rest of human genes		18,838	9,429
Healthy driver genes	All healthy drivers	Widjewadharne et al., 2020	79	64
	Only healthy drivers		36	25
	Healthy and cancer drivers		43	39
	Rest of human genes		19,470	10,206
Mendelian disease genes	All	OMIM database (McKusick-Nathans Institute of Genetic Medicine, 2020)	3568	2,615
	Recessive		1982	1,436
	Dominant		1022	757
	Rest of human genes		15,981	5,526
Total		**Unique gene loci**	**19,549**	**10,334**

Both core and context dependent essential genes had significantly higher EP scores than the rest of human genes (Figure 4-4F). On a pathway level, the median EP score correlated with the percentage of core essential genes (Figure

4-4G). By definition, the damage of core essential genes is not tolerated by cells, therefore these results agreed with the lower tolerance of high scoring pathways to damaging alterations in the germline, cancer samples and cell lines (Figure 4-4A-C). Of note, the circadian clock and chromatin organization pathways were tolerant to somatic alterations in cancer samples and cell lines, and contained few core essential genes despite their high EP score (Figure 4-4B,C,G). Context dependent essential genes showed a similar correlation as core essential genes (Figure 4-4H). Similar to core essential genes, we confirmed the circadian clock pathway as an outlier with a low percentage of context dependent essential genes. In contrast, the chromatin organization pathway contained the highest percentage of context dependent essential genes while it contained the second lowest percentage of core essential genes (Figure 4-4G-H). These results may reflect the tissue and cell-type-specific roles played by chromatin modifiers, such as regulation of gene expression in specific tissues. They also point towards a dependence of chromatin modifier essentiality on the genetic background of the cell line.

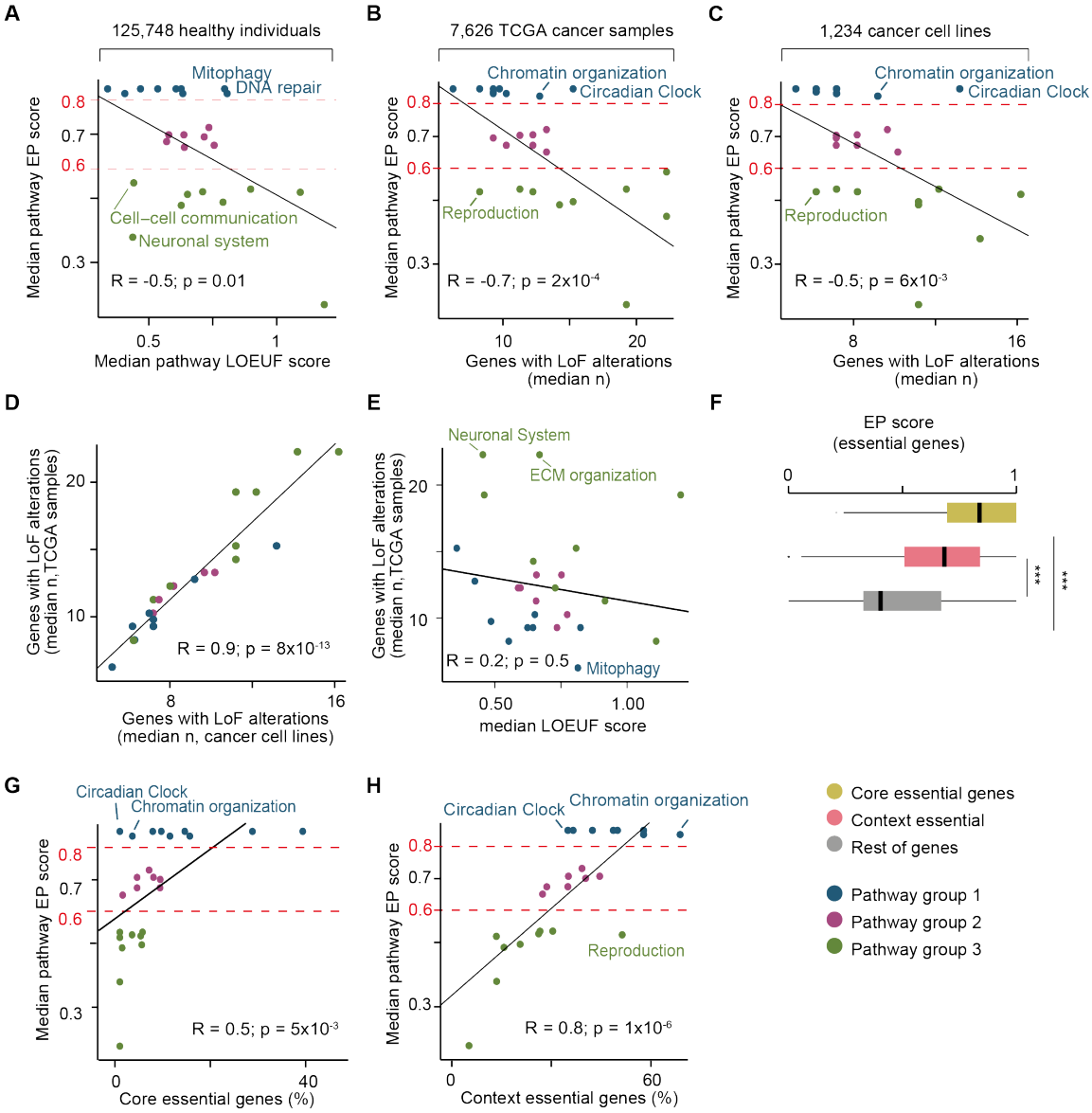


Figure 4-4 Correlation between pathway EP score and loss-of-function alterations. **A)** Correlation between the median pathway EP score and the median LOEUF score. The LOEUF score represents the loss of function observed divided by the loss of function expected upper bound fraction as obtained from whole exome sequencing data from 125,748 individuals. Low LOEUF scores indicate fewer loss-of-function alterations than expected, and therefore a high selective pressure against loss-of-function alterations. **B-C)** Correlation between the pathway median EP score and pathway median number of loss-of-function alterations per gene in B) TCGA samples or C) cancer cell lines. Briefly, for each gene, the number of samples or cell lines with a loss-of-function alteration (homozygous deletion, damaging single nucleotide variant, indel or truncation) is

counted. For each pathway, the median number of samples with loss-of-function alterations of all its member genes is shown. **D)** Correlation between the pathway median number of loss-of-function alterations per gene in TCGA samples and cancer cell lines. **E)** Correlation between the median LOEUF score and pathway median number of loss-of-function alterations per gene in TCGA. **F)** EP score distribution of essential genes. Core essential (essential in >80% of tested cell lines), context dependent essential (essential in >1 cell line and <80% of tested cell lines) and nonessential genes were compared. p-values were calculated using a two-sided Wilcoxon rank sum test. **G-H)** Correlation between the median pathway EP score and the percentage of G) core essential and H) context dependent essential genes per pathway. A, B, C, D, E, G, H) Each dot represents one functional pathway and is coloured according to its pathway group (as in Figure 4-2). Correlations were calculated using the Pearson correlation coefficient R.

4.5 Pathway EP scores predict gene involvement in disease

Cancer and Mendelian diseases are determined by the somatic or germline alteration of certain genes. Given the relationship between the EP score and tolerance towards genetic alterations, we investigated the relationship between pathway EP score and involvement in diseases. First, we investigated the connection of the EP score with cancer, using 711 cancer drivers annotated in NCG6 (Chapter 3.2.1) (Repana *et al.*, 2019) that included 239 oncogenes and 239 tumour suppressors (Table 4-2). Cancer drivers had higher EP scores than the rest of human genes (Figure 4-5A), reflecting their ubiquitous RNA and protein expression, numerous miRNA interactions, participation in complexes by encoded proteins and a central, connected and inter-connected position in the protein-protein interaction network (Chapter 3.2.2, Chapter 3.3.2) (Repana *et al.*, 2019). Similarly, tumour suppressors had higher EP scores than oncogenes (Figure 4-5A). Group 1 pathways which had high EP scores contained more cancer genes than group 2 and 3, with the exception of the ‘mitophagy’ and ‘metabolism of RNA’ pathways (Figure 4-5B). In addition, the median pathway

EP score significantly correlated with the proportion of tumour suppressors (Figure 4-5C), but not oncogenes (Figure 4-5D). Their differing EP score and participation in pathways emphasized that oncogenes and tumour suppressors are two distinct classes of cancer drivers.

As described in Chapter 3.3, several healthy drivers were identified recently (Anglesio *et al.*, 2017; Brunner *et al.*, 2019; Lac *et al.*, 2019; Lac *et al.*, 2018; Lawson *et al.*, 2020; Lee-Six *et al.*, 2019; Martincorena *et al.*, 2018; Martincorena *et al.*, 2015; Moore *et al.*, 2020; Olafsson *et al.*, 2020; Suda *et al.*, 2018; Yokoyama *et al.*, 2019; Zhu *et al.*, 2019). We were interested to which extent these genes were similar to genes altered somatically in cancer. Therefore, we performed a similar analysis of EP score and pathway distribution for 79 healthy drivers, 43 of which were also canonical cancer drivers (Wijewardhane *et al.*, 2020). Of note, we added 16 healthy drivers to the analysis described in Chapter 3.3 at a later time point. Similar to cancer genes, healthy drivers that were also cancer genes had high EP scores, whereas healthy drivers that were not cancer genes did not differ from the rest of human genes (Figure 4-5E). These results are in line with individual property differences observed in Chapter 3.2.2. Similarly, the healthy, not-cancer driver distribution across pathways did not depend on pathway EP scores (Figure 4-5F) whereas healthy cancer drivers were enriched in group 1 pathways (Figure 4-5G). In conclusion, genes with similar property profiles as the rest of human genes can drive clonal expansion in non-cancer tissue.

To explore germline alteration related disease genes, we obtained a list of 3,568 Mendelian disease genes including 1,982 genes with recessive and 1,022 genes with dominant phenotypes (McKusick-Nathans Institute of Genetic Medicine, 2020) (Table 4-2). Mendelian disease genes had a significantly higher EP score than the rest of human genes (Figure 4-5H), although it was lower than the EP score of essential (Figure 4-4F) and cancer genes (Figure 4-5A). This difference was mainly driven by dominant Mendelian disease genes, which had a higher median EP score than recessive genes (Figure 4-5H). In contrast to essential and cancer genes, pathways with high proportions of Mendelian disease genes had lower EP scores (Figure 4-5I), a trend influenced by recessive Mendelian disease

genes (Figure 4-5J,K). This indicated that the disruption of basic cellular processes through somatic alterations is more tolerated compared to germline alterations.

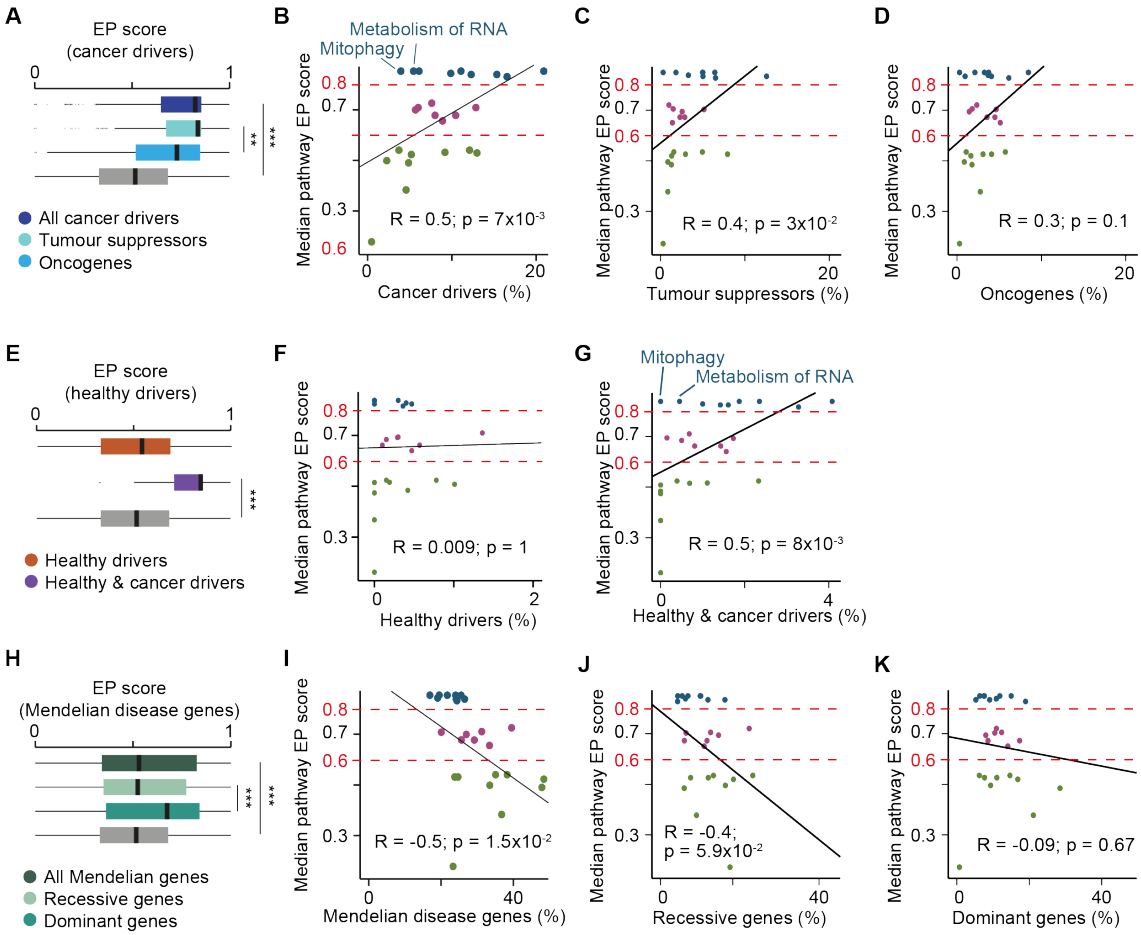


Figure 4-5 Correlation between pathway EP score and involvement in disease.

A, E, H) EP score distribution of A) cancer drivers, E) healthy drivers and H) Mendelian disease genes across pathways. p-values were calculated using a two-sided Wilcoxon rank sum test. **B-D)** Correlation of the median pathway EP score and the percentage of B) canonical cancer driver genes, C) tumour suppressors and D) oncogenes. **F-G)** Correlation of the median pathway EP score and the percentage of F) healthy drivers that are not cancer drivers and G) healthy drivers that are also cancer drivers. **I-K)** Correlation of the median pathway EP score and the percentage of I) all Mendelian disease genes, J) recessive Mendelian disease genes and K) dominant Mendelian disease genes. B, C, D, F, G, I, J, K) Each dot represents one functional pathway and is coloured according to its pathway group (as in Figure 4-2). The correlation was calculated using the Pearson correlation coefficient R.

In addition to the 711 canonical cancer drivers, we identified 1,661 candidate cancer drivers in NCG6 (Chapter 3.2.1) (Repana *et al.*, 2019). Given that candidates identified in at least two cancer sequencing screens exhibited evolutionary properties more similar to canonical cancer genes (Figure 3-2), we hypothesized that evolutionary properties could identify which candidates were likely to be true cancer drivers. This would help to prioritize cancer drivers for experimental validation. We performed a PCA as described before (Chapter 4.3) using only canonical and candidate cancer drivers. While canonical cancer genes were condensed in one area of the PCA plot (Figure 4-6A), candidates separated into two groups (Figure 4-6B). One of these had high EP scores and overlapped with canonical cancer drivers in the PCA plot. This group most likely contained true positive cancer drivers. In contrast, the other group had lower EP scores, was distant from canonical drivers in the PCA plot and was potentially enriched in false positives.

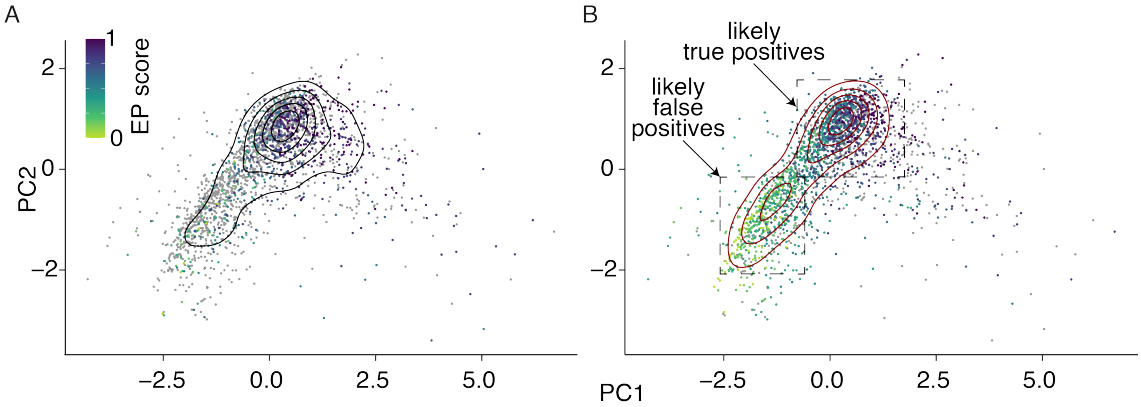


Figure 4-6 Identifying likely true positive candidate cancer genes using evolutionary properties.

PCA including nine evolutionary properties and 2,372 cancer genes. The plot shows contributions of the first and second principal components, each dot represents one gene. Genes belonging to **A)** 711 canonical cancer genes and **B)** 1,661 candidate cancer genes are indicated with colour according to their EP score. Genes which are likely true positive cancer drivers and likely false positive cancer drivers are indicated in boxes.

4.6 Conclusion

The comprehensive and integrated analysis of nine evolutionary properties enabled a broader understanding of gene evolution and function as compared to studies analysing each contribution individually. In particular, it allowed us to directly quantify the relationship between gene evolutionary properties and tolerance to perturbation, specifically in disease (Figure 4-7).

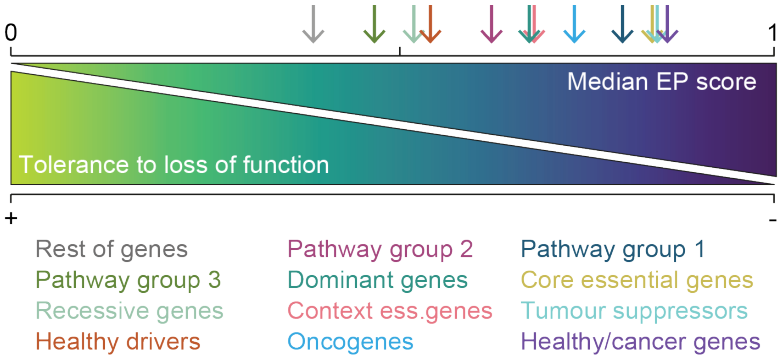


Figure 4-7 Functional and disease gene groups ordered by their median EP score. The relative position along the EP score axis is indicated with an arrow for each functional and disease gene group. High tolerance to loss-of-function alterations is indicated with a (+), low tolerance with a (-).

The high number of comparisons of evolutionary properties and pathways was hindering a clear and comprehensive analysis of the interplay of evolutionary properties with gene function. Integrating nine properties into one EP score allowed us to group pathways based on their combined properties. This revealed common functions within three distinct groups. Using the EP score, we showed a clear trend of negative selection against loss-of-function alterations in high scoring pathways. Similarly, perturbations of high scoring pathways were more likely to result in genetic diseases, confirming the central role that these pathways play in maintaining healthy, properly functioning cells.

While the median EP score was a good overall estimate of a pathway's evolutionary properties and a good indicator of its function or involvement in disease, it was not able to identify different gene groups within a pathway. Therefore, the PCA of nine evolutionary properties was a useful tool to visualize

intra-pathway heterogeneity. For example, within the reproduction pathway, two distinct gene groups in the PCA corresponded to two distinct functional groups. Similarly, the PCA was possibly able to divide candidate cancer genes into groups of false positive and true positive cancer driver genes.

Genes involved in chromatin organization had high EP scores, indicating their central role in the cell. While they were depleted in core essential genes, they contained the highest percentage of context dependent essential genes among all pathways. A possible explanation may be that the essentiality of these genes is dependent on the genetic background of the cell line. Due to synthetic lethal interaction, a gene may only be essential in cell lines that lost or downregulated its functional compensator. Taken together, these observations suggest that the chromatin organization pathway might be enriched in synthetic lethal interactors. It also has the third highest percentage of tumour suppressors (Table 7-1), which are traditionally more difficult to target therapeutically, but may be susceptible to synthetic lethality based therapy (Brunen and Bernards, 2017; Rehman *et al.*, 2010). In summary, genes involved in chromatin organization may represent valuable new targets for synthetic lethality-based cancer therapy approaches.

Chapter 5. Synthetic lethality between Epigenetic Modifiers as targetable cancer vulnerabilities

5.1 Motivation

Synthetic lethality represents a promising opportunity for cancer therapy, especially for lost tumour suppressors which cannot be targeted directly (Brunen and Bernards, 2017; Rehman *et al.*, 2010). Several approaches, including pathway synthetic lethality, synthetic dosage lethality and paralog synthetic lethality (Chapter 1.3.4), are under investigation in clinical trials. Here, we will focus on identifying additional synthetic lethal interactions between paralogs deriving from functional compensation, which may have clinical relevance for cancer therapy. We will put a special emphasis on interactions between epigenetics related genes since epigenetic modifiers are frequently lost in cancer (Feinberg *et al.*, 2016) and are difficult to target if they are tumour suppressors. In addition, as discussed in Chapter 4.6, genes involved in chromatin organization may represent interesting new targets for synthetic lethality-based cancer therapy. Examples of synthetic lethal interactions between epigenetic modifier paralogous genes have been described in several epigenetics-related functions, including *ARID1A*/*ARID1B* (Helming *et al.*, 2014) and *SMARCA2*/*SMARCA4* (D'Antonio *et al.*, 2013; Hoffman *et al.*, 2014; Oike *et al.*, 2013) which are part of the SWI/SNF chromatin remodelling complex (Mashtalir *et al.*, 2018), cohesin subunits *STAG1*/*STAG2* (Benedetti *et al.*, 2017; van der Lelij *et al.*, 2017), histone methyltransferases *EZH1*/*EZH2* (Honma *et al.*, 2017) and *KMT2A*/*KMT2B* (Ernst *et al.*, 2016) and histone acetyltransferases *CREBBP*/*EP300* (Ogiwara *et al.*, 2016). We therefore investigated whether epigenetic modifiers are enriched in paralog synthetic lethal interactions and whether evolutionary properties of epigenetic modifiers predispose them to functional compensation and synthetic lethality.

5.2 Workflow for prediction and validation of functional compensators of epigenetic modifiers in cancer genomes

To predict new paralog synthetic lethal interactions between epigenetic modifiers, we devised a multi-step analytical pipeline. First, we prioritised those epigenetic modifiers that were lost most frequently across cancer types because they had the highest relevance for potential future therapeutic application. We then predicted their potential functional compensators using sequence conservation as an indication of paralogy. The final step was to manually review literature evidence of similar function (Figure 5-1). This section focuses in detail on the identification of potential functional compensator pairs while the validation of synthetic lethality in cell lines is addressed in the next section.

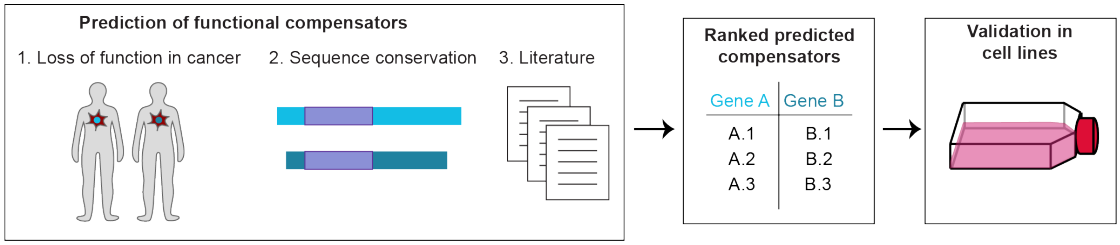


Figure 5-1 Workflow for prediction and validation of functional compensators of epigenetic modifiers in cancer genomes.

The identification of new functional compensators was performed in three steps: first, a set of candidate genes was identified based on frequent acquisition of loss-of-function alterations in TCGA. Second, pairs were identified through genetic sequence conservation. Third, potential compensator pairs were filtered based on literature evidence of similar function. Synthetic lethal interactions between the top candidates were validated in cell lines.

5.3 Prediction of functional compensators of epigenetic modifiers in cancer genomes

For the investigation of functional compensation between epigenetic modifiers, we first needed a list of genes involved in epigenetics. A consensus definition of epigenetic modifiers is challenging. For example, only the functional subunit of complexes might be considered as involved in epigenetics, or all subunits which form the complex. In addition, histone proteins are frequently modified in epigenetic processes, but do not actively contribute to the epigenetic modification process. For the purpose of this thesis, we combined three independent lists of epigenetic modifiers from the literature (Hoffman *et al.*, 2014; Medvedeva *et al.*, 2015; Plass *et al.*, 2013) and an unpublished list obtained from the Cancer Epigenetics Laboratory at the Francis Crick Institute led by Paola Scaffidi. We then investigated the involvement in epigenetics of each gene through a literature review. Based on previous synthetic lethal pairs being non-catalytical subunits of epigenetic complexes (for example, *ARID1A*/*ARID1B*), we added all subunits of complexes involved in epigenetics to the list. We excluded histone proteins as they are modified, but not modifiers themselves. Of note, eight genes involved in epigenetic modification were excluded due to their multiple functions besides epigenetics: *SALL1*, *RB1*, *SMAD4*, *TP53*, *VHL*, *MTOR*, *BRCA2* and *ATM*. Through this approach, we generated a curated list of 881 epigenetic modifiers that were categorised into histone modifiers, DNA modifiers and chromatin modifiers (Figure 5-2A, Table 7-8). Of the three categories, histone modifiers represented the largest group with 685 genes, including writers, readers, and erasers of histone acetylation, biotinylation, butyrylation, citrullination, crotonylation, glycosylation, methylation, PARylation and phosphorylation. The second largest group were 279 chromatin modifiers, including histone chaperones and proteins facilitating chromatin conformation change. Finally, DNA modifiers were the smallest group, including 51 writers, readers and erasers of DNA carboxylation, formylation, hydroxymethylation and methylation.

Of the 881 epigenetic modifiers, 322 were shared among at least three of the four sources, while 65 were added to the list based on literature curation (Figure 5-2B).

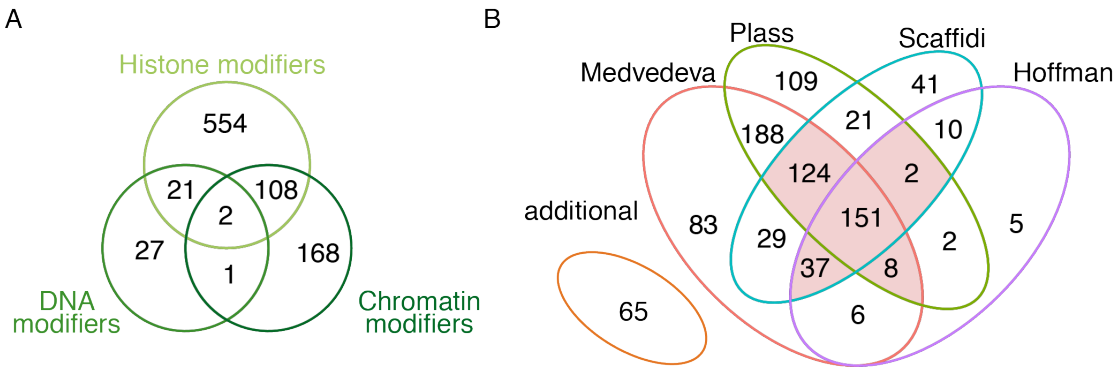


Figure 5-2 Identification of a curated list of Epigenetic Modifiers.
A) Genes identified as epigenetic modifiers were involved in histone modification, DNA modification and chromatin modification. **B)** Genes identified as epigenetic modifiers were collected from four different sources and through additional literature search.

The loss-of-function alterations in 7,828 tumour samples from 31 tumour types were annotated from TCGA data by my colleague Thanos Mourikis (Grossman *et al.*, 2016; TCGA Research Network, 2021)(Methods, Table 7-2). We defined loss-of-function alterations as homozygous deletions, or heterozygous deletions with an additional damaging mutation on the other allele.
Out of 881 epigenetic modifiers, 444 epigenetic modifiers were lost in at least one sample, and 24 epigenetic modifiers were lost in at least ten samples (Figure 5-3, Table 5-1).

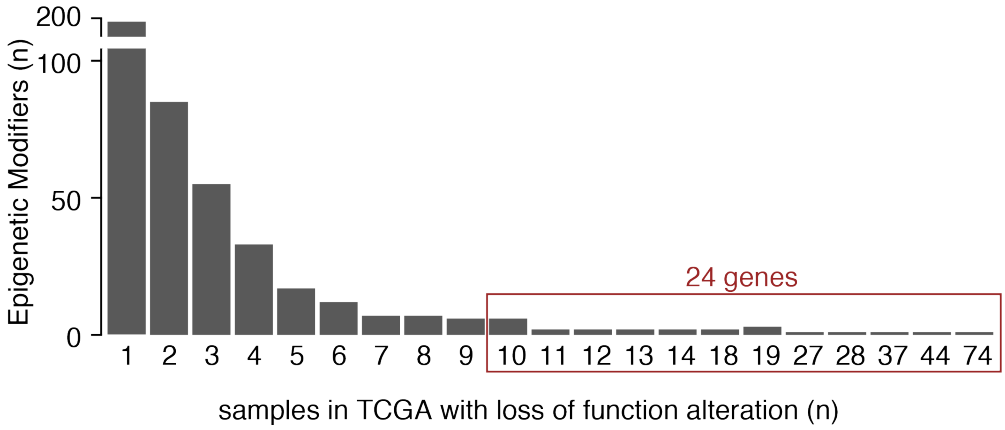


Figure 5-3 Frequency of loss-of-function alterations observed in TCGA samples for epigenetic modifiers.

The number of samples with a loss-of-function alteration in 881 epigenetic modifiers was counted in 7,828 TCGA samples. Twenty-four genes with a loss-of-function alteration in at least ten samples are highlighted.

Among these 24 genes, we identified six genes with known synthetic lethal partners (Table 5-1). Ten genes had exactly one coding duplicate gene with at least 5% sequence conservation, as determined by their protein sequence alignment to the human genome (Methods). One of these gene pairs, *CTCF*/*CTCFL,* was excluded as the genes did not compensate for each other's loss in mice (Sleutels et al., 2012), resulting in nine promising gene pairs for further investigation (Table 5-1). Of note, seven of the nine shortlisted epigenetic modifier pairs were exchangeable subunits of the same complex.

Table 5-1 Predicted gene pairs for further investigation.

Gene A indicates 24 epigenetic modifiers that were lost in at least ten TCGA cancer samples. They are listed with the number of samples in which a loss-of-function (LoF) alteration was observed. The presence or absence of exactly one coding duplicate gene with >5% coverage of the protein sequence aligned to the human genome is indicated (gene B). * highlights known synthetic lethal interactors. Based on this information, nine epigenetic modifier pairs were included for further validation.

gene symbol (gene A)	samples with LoF alterations (n)	duplicate (gene B)	included for validation
PBRM1	74	none	no
BAP1	44	none	no
MLLT3	37	*MLLT1*	yes
SETD2	28	none	no
ARID1A	27	*ARID1B**	no
KMT2A	19	*KMT2B**	no
NSD1	19	>1 duplicate	no
NCOR1	19	*NCOR2*	yes
KMT2C	18	*KMT2D*	yes
MYOCD	18	>1 duplicate	no
SMARCA2	14	*SMARCA4**	no
SMARCA4	14	*SMARCA2**	no
CTCF	13	*CTCFL*	no: excluded based on literature evidence
ARID1B	13	*ARID1A**	no
CREBBP	12	*P300**	no
BRD7	12	*BRD9*	yes
TFDP1	11	*TFDP2*	yes
KDM6A	11	*KDM6B*	yes
CHD1	10	*CHD2*	yes
SPEN	10	none	no
HCFC1	10	*HCFC2*	yes
TBL1X	10	*TBL1XR1*	yes
CHD9	10	>1 duplicate	no
HASPIN	10	none	no

5.4 Validation of functional compensators of epigenetic modifiers in cancer genomes

5.4.1 Screening of nine gene pairs

The computational pipeline predicted nine epigenetic modifier pairs as potential functional compensators. To investigate these gene pairs, we collaborated with the Crick High Throughput Screening Science Technology Platform to set up a screen using siRNA knockdown and CRISPR Cas9 knockout in cell lines. In this screen, we compared cell growth between single gene and double gene knockdown/knockout for each of the nine gene pairs (Figure 5-4). As a positive control, we included the gene pair *EZH1*/*EZH2* (Honma *et al.*, 2017). We used Hacat, an immortalized human adult keratinocyte cell line, and HEK293, an immortalized human embryonic kidney cell line. Both cell lines did not have any alterations in the 20 investigated genes, as reported in DepMap (Chapter 2.2.3). For each gene pair, we knocked down gene A, gene B or both genes simultaneously using an siRNA pool library (Methods). For CRISPR Cas9 knockout experiments, we applied a similar approach, transfecting gRNAs into Hacat and HEK293 cells with stable Cas9 expression.

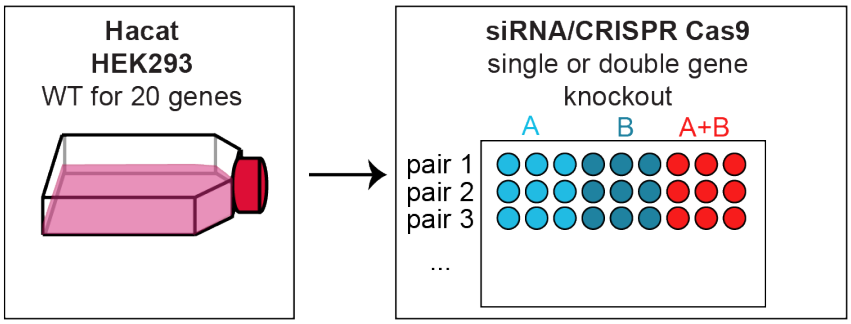


Figure 5-4 Investigation of synthetic lethality between nine epigenetic modifier pairs.

Hacat and HEK293 cells were transfected with siRNA or gRNA using Lipofectamine™2000 (Hacat) or INTERFERin® (HEK293) transfection reagents. For each gene pair, the genes were knocked down/out individually and as a pair. Cells were seeded in 96-well plates and growth was observed for up to 6 days using an Incucyte® ZOOM system.

Firstly, we optimized cell seeding density in 96-well plates and transfection reagents for both cell lines. Transfection with the negative control (non target control (nTC) for CRISPR Cas9 knockout and RISCFree for siRNA knockdown) resulted in exponential growth, thereby confirming the successful setup of the screening assays (Figure 5-5A-D, Figure 7-2). Next, we confirmed cell death upon treatment with the positive controls, *PLK1* for CRISPR Cas9 knockout and *UBB* for siRNA knockdown. These control genes were previously confirmed to reduce cell growth in proliferation assays performed by the Crick High Throughput Screening Science Technology Platform. For the CRISPR Cas9 positive control knockout, we observed that HEK293 cells were able to reach confluency, albeit more slowly compared to the negative control, while Hacat cells did not reach confluency (Figure 5-5A,C). The final effect of siRNA positive control was comparable in both cell lines, but HEK293 cells initially grew for approximately 24 hours before dying (Figure 5-5B,D).

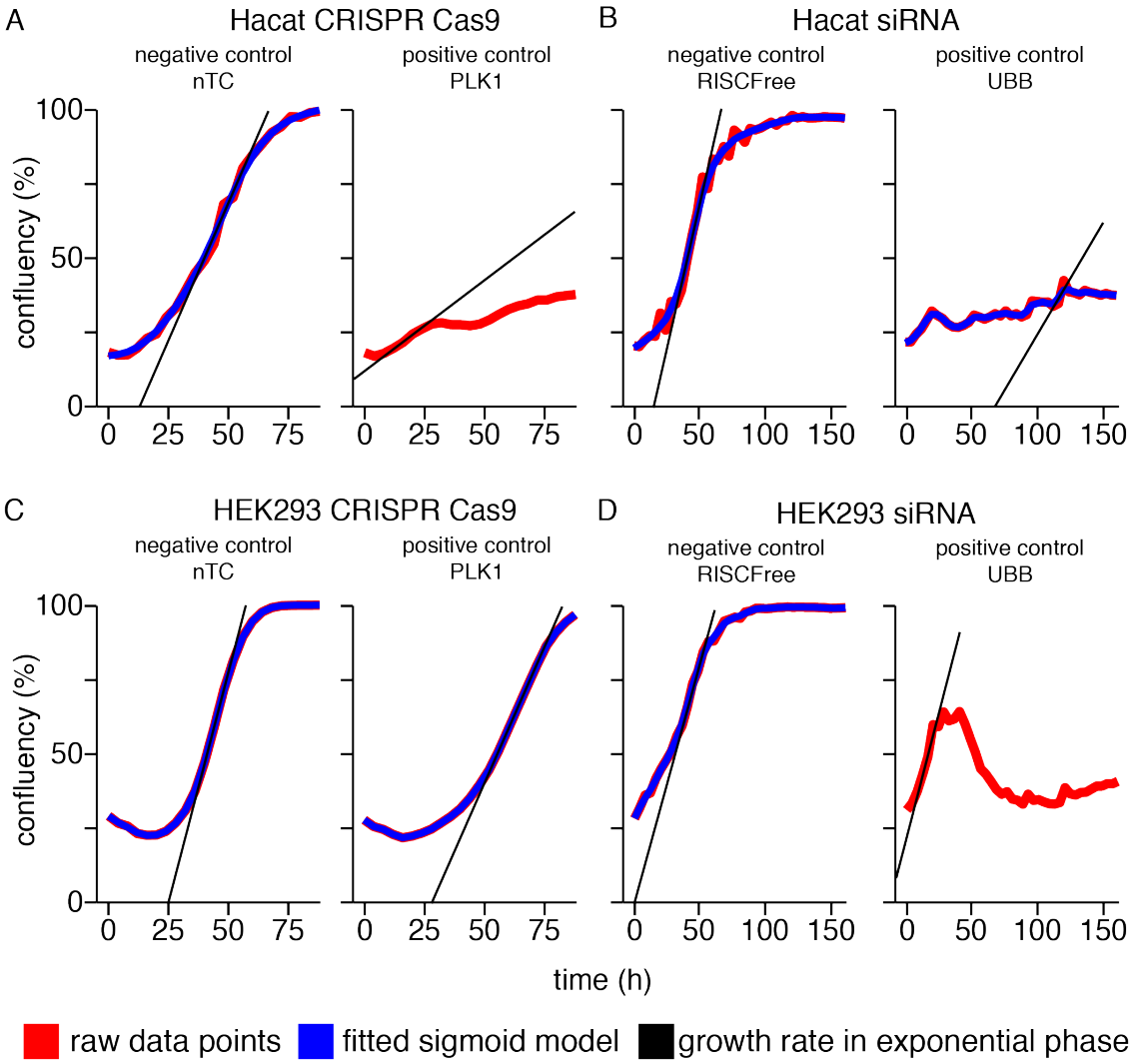


Figure 5-5 Positive and negative control of siRNA knockdown and CRISPR Cas9 knockout in Hacat and HEK293 cells.

A) CRISPR Cas9 knockout of nTC and *PLK1* in Hacat cells. **B)** siRNA knockdown of RISCFree and *UBB* in Hacat cells. **C)** CRISPR Cas9 knockout of nTC and *PLK1* in HEK293 cells. **D)** siRNA knockdown of RISCFree and *UBB* in HEK293 cells. Briefly, cells were seeded at 8000 cells/well (Hacat) and 6000 cells/well (HEK293) into 96 well plates. Hacat cells were transfected using 0.2µl/well INTERFERin®, HEK293 cells were transfected using 0.2 µl/well Lipofectamine™2000 reagent. Growth was assessed using an Incucyte® ZOOM system. A fitted sigmoid model and linear model of the growth rate in the exponential phase were calculated using the gcFitSpline function from the R grofit package (Kahm *et al.*, 2010). One representative curve out of three replicates is shown.

We did not observe cell death upon single gene or double gene knockout/ knockdown in any of the tested gene pairs, including the positive control pair *EZH1*/*EZH2* (Figure 7-2). Only for the simultaneous CRISPR Cas9 knockout of *NCOR1* and *NCOR2* in HEK293, we observed a significant decrease in cell growth compared to single gene knockout (Figure 7-2C). This effect was not observed in siRNA knockdown, or Hacat cells. Compared to single gene *TBL1X* or *TBL1XR1* knockdowns, Hacat cells did not reach full confluency when both *TBL1X* and *TBL1XR1* were knocked down (Figure 5-6A). These results indicated a potential functional compensation mechanism between the two genes. Notably, this effect was not observed in Hacat cells using CRISPR Cas9 knockout (Figure 5-6B), or using either approach in HEK293 cells (Figure 5-6C,D).

The variable results between siRNA and CRISPR Cas9 knockout, as well as the lack of synthetic lethality between the *EZH1*/*EZH2* control pair, suggested a high proportion of false negative results in the screen. This was possibly due to the short time over which we conducted the screen. While siRNA knockdown directly decreases the RNA levels inside a cell, a CRIPSR Cas9 knockout on DNA level may take some time to result in decreased RNA expression and protein levels. Therefore, it is unclear why we only observed an effect for double knockout of *NCOR1*/*NCOR2* in the fast-growing HEK293 cell line and using CRISPR Cas9 knockout. In contrast, we observed a limited negative effect of *TBL1X*/*TBL1XR1* double knockdown on cell growth towards the end of the experiment, but only using siRNA knockdown, and only in the slower growing Hacat cells. Since we were only able to observe cell growth for a few days after knockdown/ knockout before cells reached confluency, the time frame to observe an effect was possibly too short. For this reason, we further investigated the effect of gene pair knockout over a longer time frame via a single gene pair validation approach.

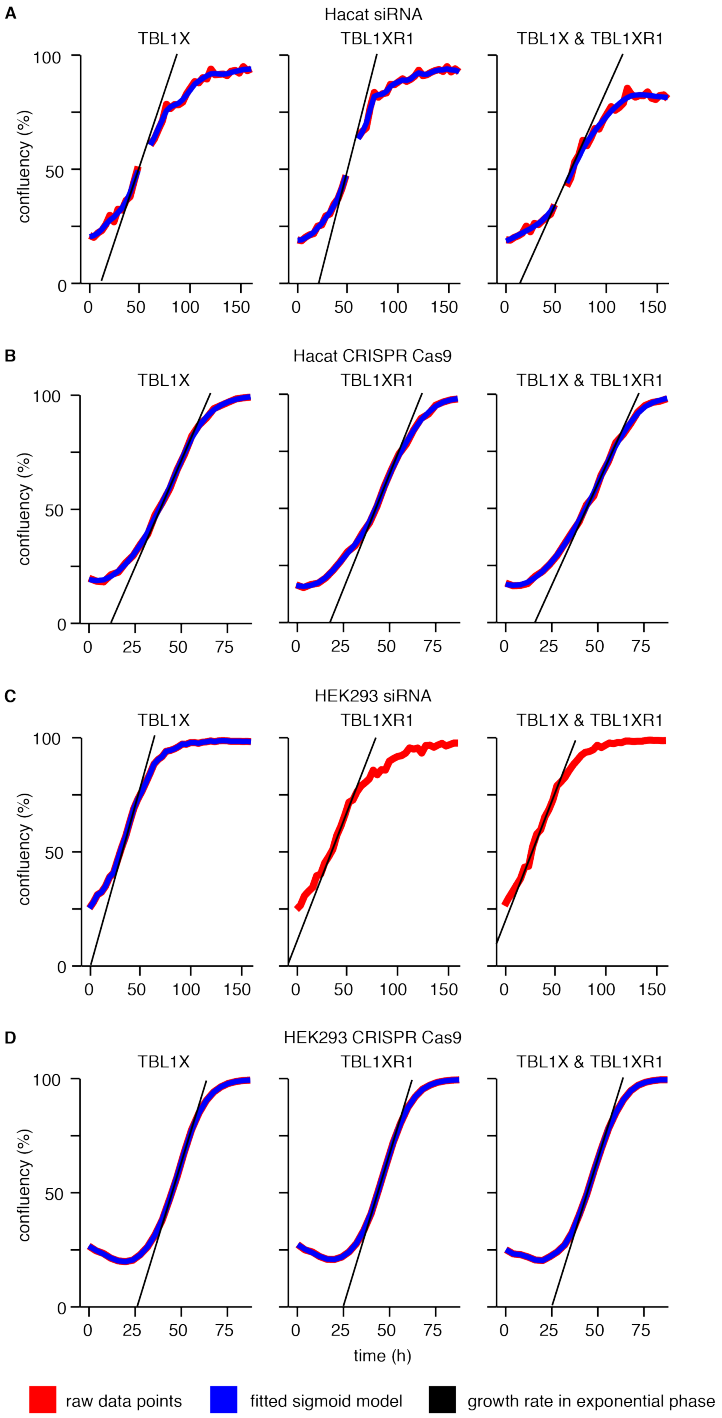


Figure 5-6 Proliferation after siRNA knockdown and CRISPR Cas9 knockout of *TBL1X*, *TBL1XR1* or both genes in Hacat and HEK293 cells.

Cell proliferation was observed after **A)** siRNA knockdown of *TBL1X*, *TBL1XR1* or both genes in Hacat cells. **B)** CRISPR Cas9 knockout of *TBL1X*, *TBL1XR1* or both genes in Hacat cells. **C)** siRNA knockdown of *TBL1X*, *TBL1XR1* or both genes in HEK293 cells. **D)** CRISPR Cas9 knockout of *TBL1X*, *TBL1XR1* or both genes in HEK293 cells. Briefly, cells were seeded at 8000 cells/well (Hacat) and

6000 cells/well (HEK293) into 96 well plates. Hacat cells were transfected using 0.2µl/well INTERFERin®, HEK293 cells were transfected using 0.2µl/well Lipofectamine™2000. Growth was assessed using an Incucyte® ZOOM system. A fitted sigmoid model and linear model of the growth rate in the exponential phase was calculated using the gcFitSpline function from the R grofit package (Kahm *et al.*, 2010). One representative curve out of three replicates is shown.

5.4.2 Single gene pair validation approach

TBL1X/*TBL1XR1* double knockdown in Hacat cells pointed towards a negative functional interaction between these paralogs. *MLLT3* was the epigenetic modifier most frequently altered in TCGA samples (Table 5-1). We designed a CRIPSR Cas9 knockout experiment that allowed us to observe cell viability at chosen time points after *TBL1X*/*TBL1XR1* and *MLLT1*/*MLLT3* knockout. We reasoned that this should allow us to identify the time point at which epigenetic changes induced by the double gene loss would show an effect on proliferation. We knocked out either one gene at a time or both genes simultaneously by introducing recombinant CRISPR Cas9 protein and synthetic gRNAs into the cell through nucleofection (Methods). Instead of directly assessing cell viability, we grew the cells in culture for 14 days before seeding them for a proliferation assay (Figure 5-7). We expected to see no difference in proliferation between control cells and cells which lost one gene. In contrast, for the loss of both genes, we expected to observe either of three conditions: 1) same proliferation levels for gene pairs with no genetic interaction, 2) reduced proliferation for synthetic sick interactions, and 3) no cell proliferation for synthetic lethal interactions (Figure 5-7). To assess knockout efficiency throughout the experiment for each gene, we used the Synthego ICE tool (Hsiau et al., 2018). This tool uses Sanger sequencing profiles to estimate the proportion of alleles in a cell pool that are affected by a damaging alteration introduced through CRISPR Cas9 editing. These include frameshift indels or indels that are larger than 21 base pairs. In contrast to time-consuming next-generation sequencing, this tool provides an alternative to evaluate knockout success in ongoing experiments.

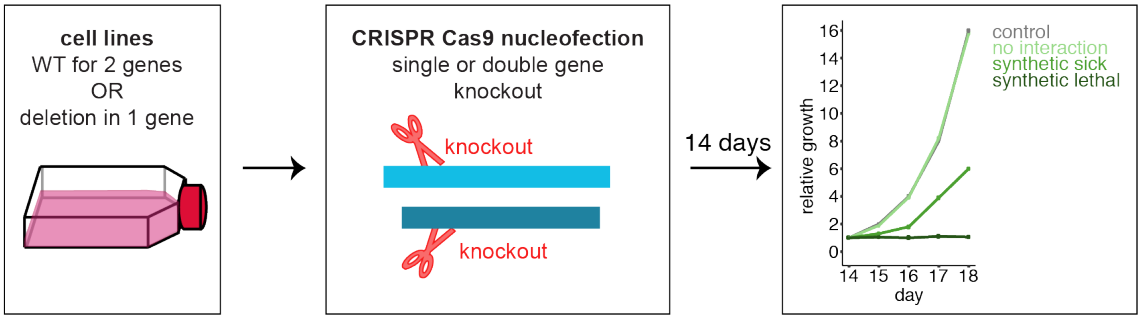


Figure 5-7 Single gene pair validation using CRISPR Cas9 knockout.
Cell lines for validation were wild type (WT) for both genes or had a homozygous deletion in one of the genes. Single gene or double gene knockout was performed using CRISPR Cas9 editing, proliferation was assessed 14 days after knockout.

5.4.3 Validation of synthetic lethality between *TBL1X*/*TBL1XR1*

Based on the screening of nine gene pairs, the gene pair *TBL1X*/*TBL1XR1* was selected for further investigation. *TBL1X* and *TBL1XR1* are interchangeable subunits of the N-CoR and SMRT complexes, which perform histone deacetylation and gene repression through their subunit HDAC3 (Guenther et al., 2000; Li et al., 2000; Yoon et al., 2003). There is approximately 75% sequence conservation between *TBL1X* and *TBL1XR1*. Both genes interact with the N-Cor complex through two conserved regions and recruit the complex to H2B and H4 histone components through an additional conserved region (Yoon *et al.*, 2003). While their individual knockdown did not affect activity of N-CoR and SMRT complexes, their combined knockdown resulted in the inability of the complexes to repress their target genes (Yoon *et al.*, 2003).

TBL1X was damaged in Bladder Urothelial Carcinoma, Ovarian serous cystadenocarcinoma, Colon adenocarcinoma, Glioblastoma multiforme, Head and Neck squamous cell carcinoma, Kidney Chromophobe and Lung squamous cell carcinoma. Therefore, we chose the model cell lines for further validation based on their alteration profile obtained from DepMap (Chapter 2.2.3). We chose NCIH1975 (lung cancer), which was wild type (WT) for both genes; NCIH2030 (lung cancer), which was WT for *TBL1XR1* and had a homozygous deletion of *TBL1X* (dTBL1X); and TEN (endometrial cancer), which was WT for

both genes. We confirmed expression of *TBL1X* and *TBL1XR1* in all cell lines (Figure 5-8). The exception was NCIH2030, which did not express *TBL1X*, confirming the homozygous deletion (Figure 5-8).

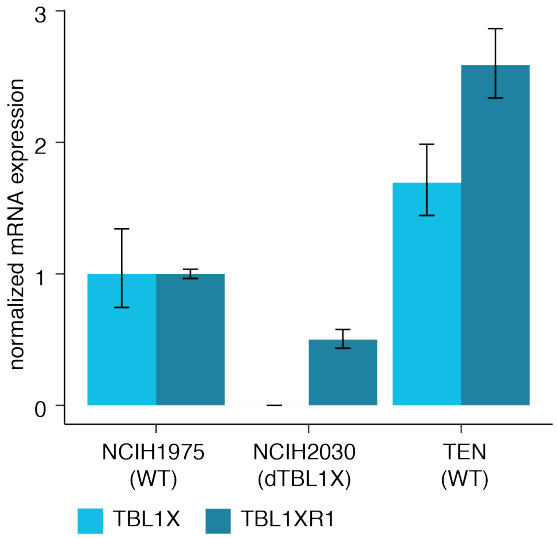


Figure 5-8 Expression of *TBL1X* and *TBL1XR1* in NCIH1975, NCIH2030 and TEN. RNA expression of *TBL1X* and *TBL1XR1* were quantified using qPCR in one biological replicate and technical triplicates. Expression levels were normalized to *GAPDH* expression levels in each cell line and normalized to expression levels of the respective gene in NCIH1975 cells using the ΔΔCt method.

We knocked out either *TBL1X* and *TBL1XR1* individually or both genes simultaneously in WT cell lines, and only *TBL1XR1* in NCIH2030. Three days after nucleofection, the knockout efficiency for each cell line was approximately 75% (Figure 5-9A-C). We only observed approximately 60% efficiency for *TBL1XR1* knockout in double knockout NCIH1975 cells (Figure 5-9B). Knockout remained stable over time with the exception of a slight drop in knockout efficiency of *TBL1XR1* in NCIH1975 and NCIH2030 after 14 days (Figure 5-9A,B). The knockout efficiency of *TBL1XR1* in double edited NCIH1975 cells increased to approximately 75% after 14 days, indicating a potentially faulty measurement of the lower result at day 3 (Figure 5-9B). Knockout efficiency after 30 days was tested only for NCIH2030 cells, and knockout was not detected (Figure 5-9A).

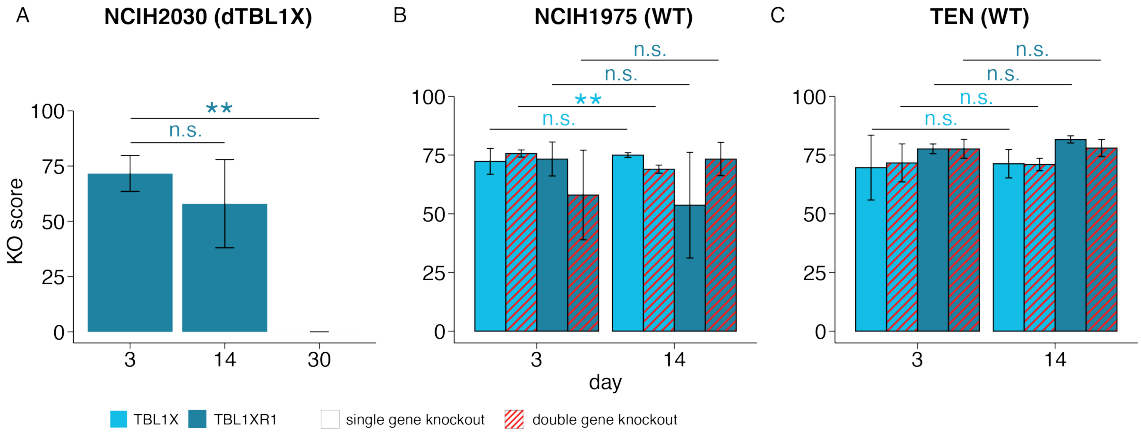


Figure 5-9 Knockout efficiency of *TBL1X*, *TBL1XR1* or double gene knockout. Recombinant Cas9 protein and sgRNAs were introduced into **A)** NCIH2030, **B)** NCIH1975 and **C)** TEN cells using nucleofection. At 3, 14 and 30 days after nucleofection, DNA was extracted from cells. Loci with expected CRISPR Cas9 editing were amplified by PCR and sequenced using Sanger sequencing. The Synthego ICE tool (Hsiau *et al.*, 2018) was used to assess the percentage of alleles of each respective gene that were altered by a damaging alteration following CRISPR Cas9 editing. KO scores between different time points were compared and p-values were calculated using a two-sided t-test. *p<0.05; **p<0.01; n.s. not significant.

After 14 days, NCIH2030 cells only reached 0.4-fold proliferation upon *TBL1XR1* knockout compared to nTC knockout (Figure 5-10A,B), indicating that the homozygous deletion of *TBL1X* is incompatible with *TBL1XR1* loss. However, we did not observe this effect for the double knockout of *TBL1X* and *TBL1XR1* in WT cell lines: neither NCIH1975 nor TEN cells exhibited a difference in proliferation after single or double gene knockouts compared to the nTC knockout (Figure 5-10C-F). After 30 days, there was no difference in proliferation between control and *TBL1XR1* knockout cells for NCIH2030, indicating that the WT cells had outcompeted the edited cells (Figure 5-10G,H). This confirmed that the knockout of *TBL1XR1* in NCIH2030 was incompatible with its homozygous deletion of *TBL1X* and resulted in a selective disadvantage.

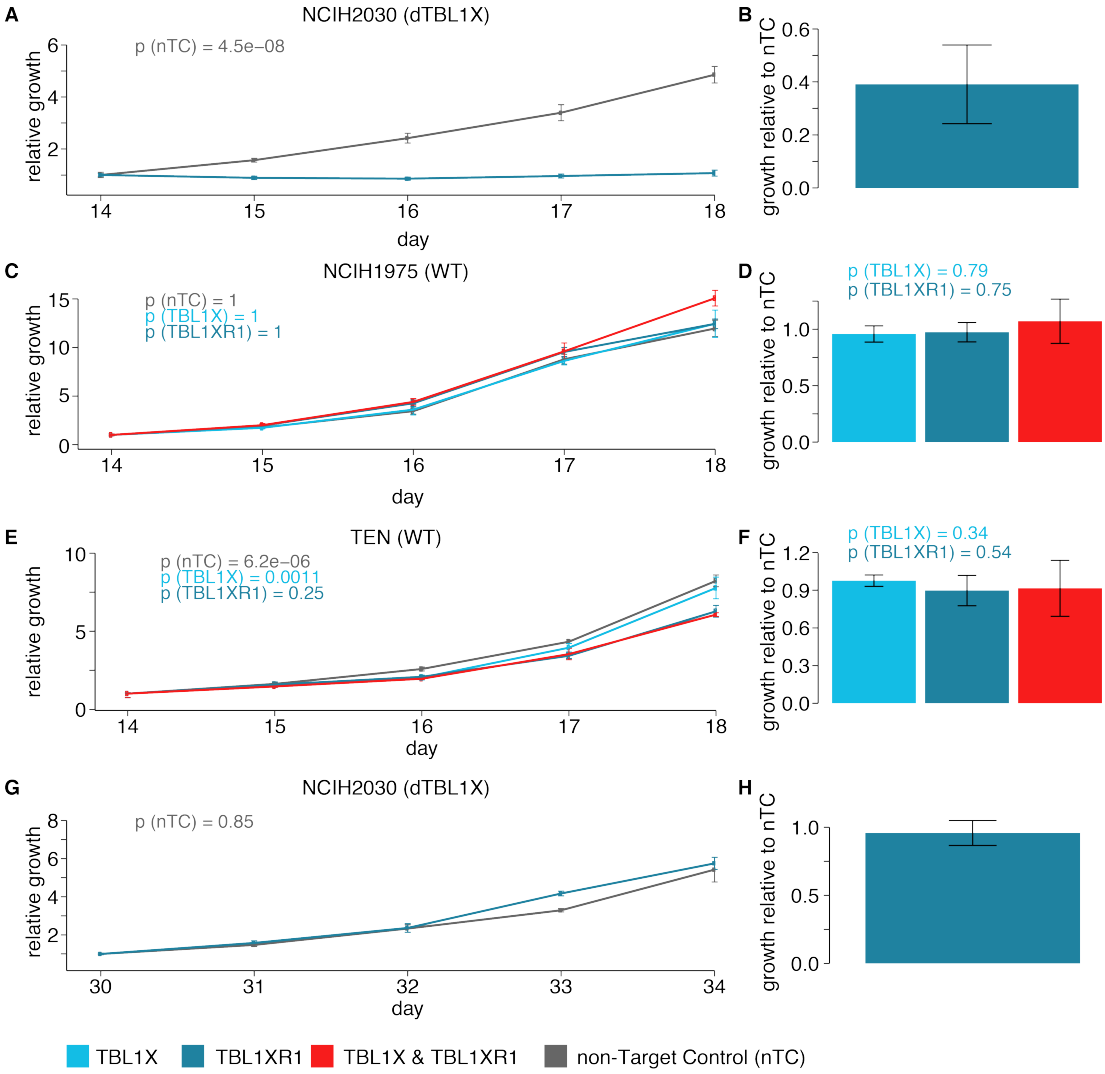


Figure 5-10 Proliferation of cell lines with *TBL1X*, *TBL1XR1* or double gene knockout

A,C,E,G) Recombinant Cas9 protein and sgRNAs were introduced into A,G) NCIH2030, C) NCIH1975 and E) TEN cells using nucleofection. After 14 or 30 days, proliferation assays were started and run for four days. A representative proliferation assay out of three replicates (Figure 7-3) is shown. P values were calculated using a one-sided t test and indicate significance at the last day of the assay between nTC and *TBL1XR1* knockout cells (A, G) or between nTC, *TBL1X* or *TBL1XR1* knockout and the double knockout (C,E). **B,D,F,H)** Relative quantification of cell proliferation compared to nTC cells at the end point. Values represent the mean of three replicates, error bars indicate +/- standard deviation intervals. P values were calculated using a one-sided t test.

TBL1XR1 RNA was still expressed upon *TBL1XR1* knockout in NCIH1975 and NCIH2030 (Figure 5-11). Since we observed a growth inhibition by *TBL1XR1* knockout in NCIH2030 cells, the RNA levels were not a good indicator of knockout success. While mRNAs with damaging base pair deletions or insertions may lead to non-functional proteins, they might still be detected by qPCR.

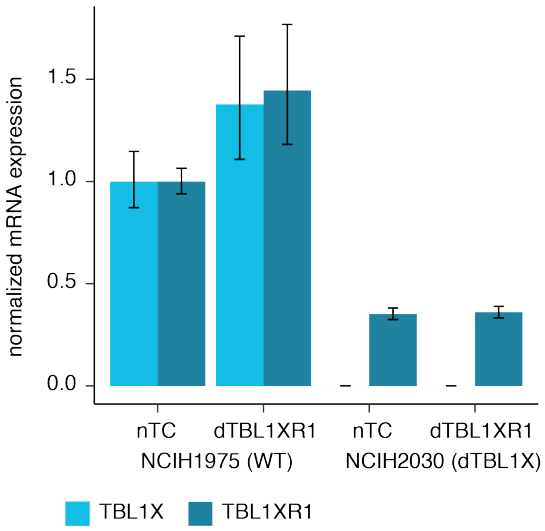


Figure 5-11 Expression levels of *TBL1X* and *TBL1XR1* in WT and edited cell lines. RNA expression of *TBL1X* and *TBL1XR1* was quantified using qPCR in one biological replicate and technical triplicates. Expression levels were normalized to *GAPDH* expression levels in each cell line using the ΔΔCt method and normalized to expression levels of the respective gene in NCIH1975 nTC edited cells.

5.4.4 Validation of synthetic lethality between *MLLT1*/*MLLT3*

The second gene pair we chose for validation was *MLLT1*/*MLLT3*, since *MLLT3* was the most frequently altered gene in TCGA samples within the group of predicted synthetic lethal pairs (Table 5-1). Even though the screen did not indicate synthetic lethality between *MLLT1*/*MLLT3*, we hypothesised that this was likely due to a false negative signal. *MLLT1* and *MLLT3* have approximately 30% sequence conservation, including a YEATS domain which binds to acetylated histones. These include histone 3 lysine 9 acetylation (H3K9ac), histone 3 lysine

18 acetylation (H3K18ac) and histone 3 lysine 79 acetylation (H3K79ac) (Li et al., 2014; Wan et al., 2017). *MLLT1* and *MLLT3* are interchangeable subunits of the DOT1L complex (Shen et al., 2013), which catalyses the methylation of histone 3 lysine 79 (H3K79) (Feng et al., 2002). Inhibition of *DOT1L* reduces H3K79 di- and tri-methylation and leads to a reduction of transcription at enhancer elements (Godfrey et al., 2019). In addition, *MLLT1* and *MLLT3* are exchangeable subunits of the Super Elongation Complex, which enhances transcription through its interaction with RNA polymerase II (He et al., 2011). In this context, knockdown of *MLLT1* results in increased expression of *MLLT3* and its incorporation into the Super Elongation Complex, but not vice versa (He *et al.*, 2011).

Thirty-seven TCGA cancer samples had a damaging alteration of *MLLT3*, including twelve glioblastoma samples (8.4% of the total glioblastoma samples in TCGA). Therefore, we investigated the potential synthetic lethal interaction of *MLLT1*/*MLLT3* in four glioblastoma cell lines. As determined using alteration and copy number data from DepMap (Chapter 2.2.3), KNS42, SF268 and HCT116 were WT for both genes, and U87MG had a homozygous deletion of *MLLT3*. Notably, both *MLLT1* and *MLLT3* were expressed most abundantly in HCT116 (Figure 5-12). In contrast, *MLLT1* expression was low in SF268 and U87MG cells. Low *MLLT3* expression was also observed in SF268, and in line with the homozygous deletion, not detectable in U87MG.

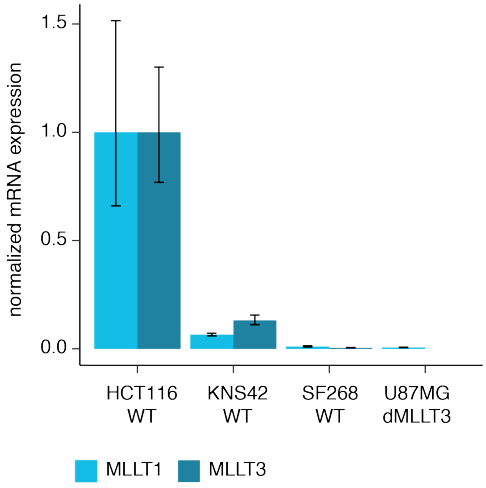


Figure 5-12 Expression levels of *MLLT1* and *MLLT3* in HCT116, KNS42, SF268 and U87MG.

RNA expression of *MLLT1* and *MLLT3* was quantified using qPCR in one biological replicate and technical triplicates. Expression levels were normalized to *GAPDH* expression levels in each cell line using the ΔΔCt method and normalized to expression levels of the respective gene in HCT116 cells.

Single gene knockout of *MLLT1* and *MLLT3* and double gene knockout was successful in all cell lines, with highest efficiency in KNS42 (Figure 5-13). We first showed that *MLLT1* knockout in WT cell lines had no effect on cell growth. We saw a decrease in proliferation of approximately 10% for KNS42 (Figure 5-14A,B), whereas no proliferation difference occurred in SF268 cells (Figure 5-14C,D). U87MG (*dMLLT3*) cells did not exhibit reproducible growth reduction upon *MLLT1* knockout (Figure 5-14E,F), indicating that a loss of both genes did not have an impact on cell fitness. In WT HCT116 cells, knockout of *MLLT1* led to a decrease in proliferation compared to control cells, but the additional knockout of *MLLT3* did not have any additional effect on proliferation (Figure 5-14G,H). Overall, these results did not support a synthetic lethal interaction between the two genes.

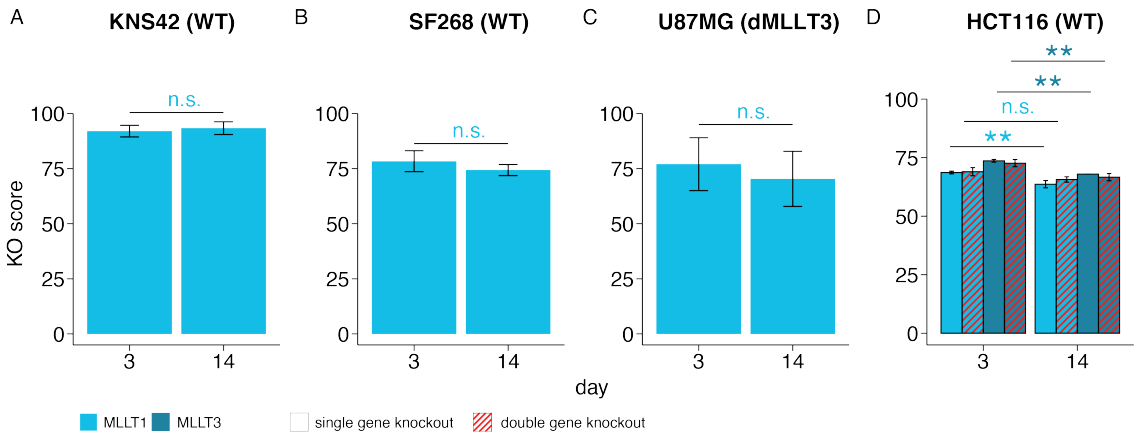


Figure 5-13 Knockout efficiency of *MLLT1*, *MLLT3* or double gene knockout. Recombinant Cas9 protein and sgRNAs were introduced into **A)** KNS42, **B)** SF268, **C)** U87MG and **D)** HCT116 cells using nucleofection. At several time points after nucleofection, DNA was extracted from cells. Loci with expected CRISPR Cas9 editing were amplified by PCR and sequenced using Sanger sequencing. The Synthego ICE tool (Hsiau *et al.*, 2018) was used to assess the percentage of alleles of each respective gene that was altered by a damaging alteration following CRISPR Cas9 editing. KO scores between different time points were compared and p-values were calculated using a two-sided t-test. *p<0.05; **p<0.01; n.s. not significant.

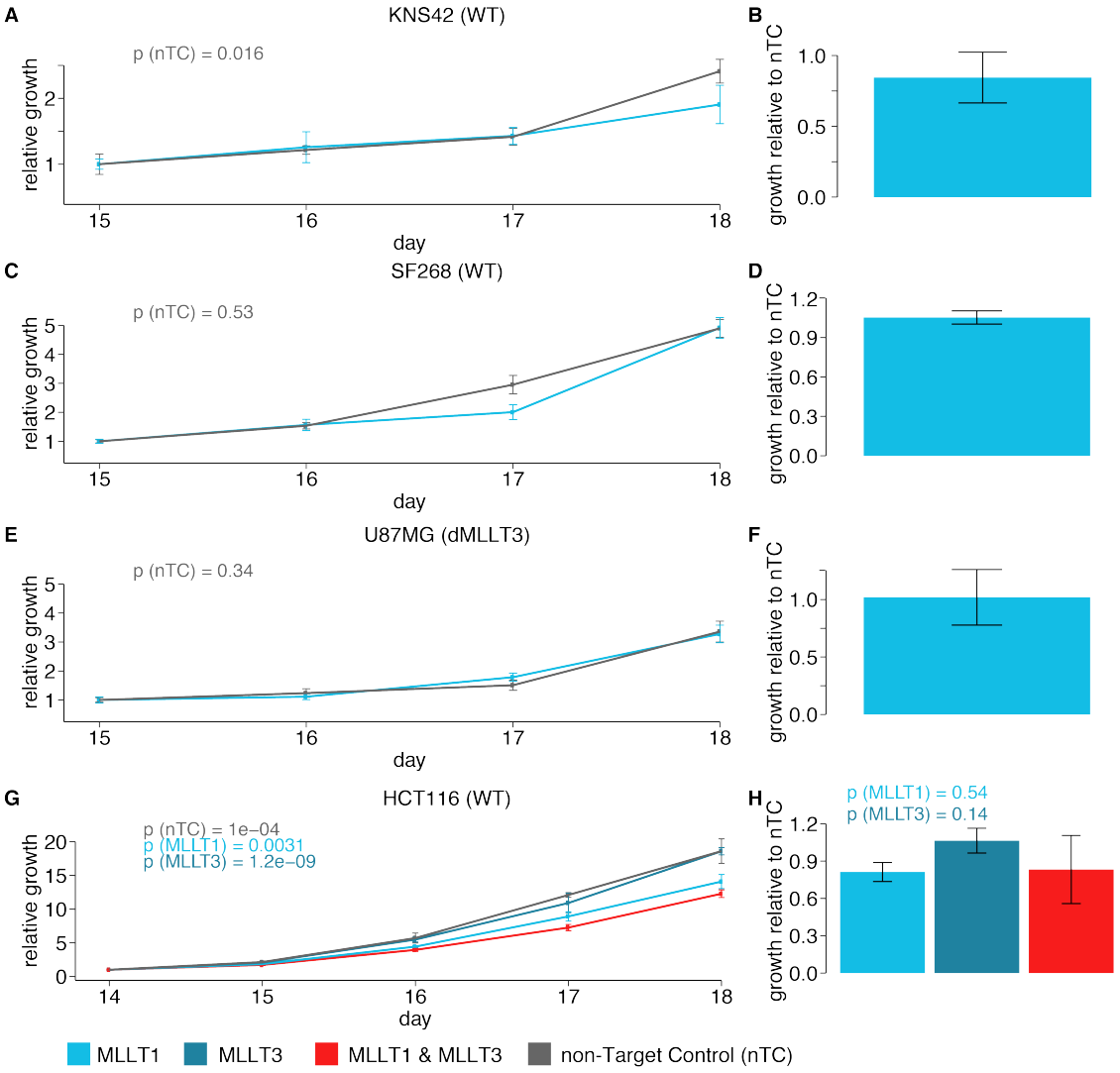


Figure 5-14 Proliferation of cell lines with *MLLT1*, *MLLT3* or double gene knockout. **A,C,E,G)** Recombinant Cas9 protein and sgRNAs were introduced into A) KNS42, C) SF268, E) U87MG and G) HCT116 cells using nucleofection. After 14 days, proliferation assays were started and run for four days. Of note, cell quantities were normalized to day one instead of day zero for A, C and E to allow for cells to attach to the plate. A representative proliferation assay out of three replicates (Figure 7-4) is shown. P values were calculated using a one-sided t test and indicate significance at the last day of the assay between nTC and *MLLT1* knockout cells (A, C, E) or between nTC, *MLLT1* or *MLLT3* knockout and the double knockout (G). **B,D,F,H)** Relative quantification of cell proliferation compared to nTC cells at the end point. Values represent the mean of three replicates, error bars indicate +/- standard deviation intervals. P values were calculated using a one-sided t test.

One of the functions of *MLLT1* and *MLLT3* is their engagement in the DOT1L complex that catalyses the methylation of H3K79 to H3K79me2 and H3K79me3, and its inhibition leads to the depletion of both methylation marks (Godfrey *et al.*, 2019). If MLLT1 or MLLT3 are necessary components of the DOT1L complex, these methylation marks should also be depleted in cells that lost both of them. However, we did not observe any differences between H3K79me2 and H3K79me3 protein levels in the cells that lost only *MLLT1* or the cells that lost both genes, as compared to WT cells (Figure 5-15). This indicated that the DOT1L complex performs its function independently of MLLT1 and MLLT3.

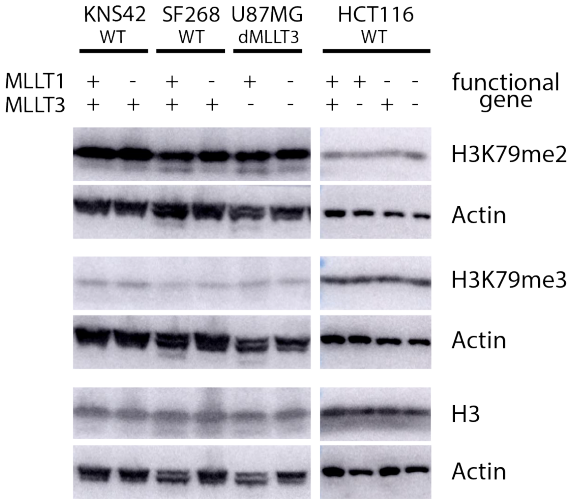


Figure 5-15 Western blot of H3K79me2 and H3K79me3 marks.
MLLT1 was knocked out in KNS42, SF268 and U87MG cells. *MLLT1*, *MLLT3* or both genes were knocked out in HCT116. Protein was extracted from cells 14 days after gene knockout and H3K79me2 and H3K79me3 marks were quantified using a Western Blot. Actin and histone H3 are shown as reference. (+) functional gene, (-) damaged gene through homozygous deletion or CRISPR Cas9 editing.

5.4.5 Conclusion

Through the integration of four lists of epigenetic modifiers and a thorough literature research, we identified 881 genes involved in epigenetic modification. Given the variable definitions of epigenetics, this list may be restricted or expanded, depending on the exact biological question. Using a computational

pipeline, we prioritized nine gene pairs that may constitute synthetic lethal partners, based on sequence conservation and frequency of loss-of-function alterations in cancer. Their validation with siRNA and CRISPR Cas9 screens led to one potentially interesting candidate pair, *TBL1X*/*TBL1XR1.* Due to the positive control not showing a signal, we could not make a clear statement about the relationship between the remaining gene pairs. We reasoned that a proliferation assay started right after knockout or knockdown was not able to capture synthetic lethal effects. DNA alterations require some time to result in a lower level of RNA and consecutively functional protein, and finally a change in the targeted epigenetic mark. Increasing the time between knockout of *TBL1X*/*TBL1XR1* and proliferation assessment indeed supported synthetic lethality between the two genes in NCIH2030, but not NCIH1975 and TEN. This is most likely explained by the context dependency of functional compensation mechanisms. In addition, we observed the lethal phenotype in the cell line with a pre-existing homozygous deletion of *TBL1X*. Therefore, the double knockout of *TBL1X* and *TBL1XR1* might be less effective than one pre-existing homozygous deletion and sufficient functional proteins may be retained to place the epigenetic mark. We were not able to show this using a qPCR, but since the qPCR only measured the transcript level, measuring the reduced amounts of functional protein after knockout might be more suitable.

The second potentially interesting candidate pair, *MLLT1*/*MLLT3*, is not a synthetic lethal pair. We investigated the effect of *MLLT1*/*MLLT3* loss on the activity of the DOT1L complex, a complex which contains either *MLLT1* or *MLLT3* as interchangeable subunits. While *DOT1L* inhibition was shown to lead to a decrease in H3K79 methylation marks (Godfrey *et al.*, 2019), the loss of both *MLLT1* and *MLLT3* did not affect H3K79 methylation. This indicates that the DOT1L complex can function independently of *MLLT1* and *MLLT3*.

This prediction pipeline only investigated epigenetic modifiers with very strictly defined damaging alterations in cancer samples. This had some implications on the analysis we were able to perform. First, while we predicted synthetic lethal epigenetic modifiers, the question whether they are enriched in potential synthetic lethal pairs remained unanswered. Second, the strict definition of damaging

alterations restricted the genes which were included in downstream analysis. Finally, the computational prediction did not consider whether the function of the gene pairs is essential to the cell. To address these issues, we improved the prediction pipeline as described in the next chapter.

5.5 Improved workflow for prediction and validation of functional compensators in cancer genomes

We developed an improved pipeline (Figure 5-16) based on updated gene properties described in Chapter 2.1.2.

First, we included all human protein-coding genes in the prediction, enabling a comparison of epigenetic modifiers with the rest of human genes. Second, we identified compensator pairs including exactly two genes through a more stringent sequence conservation threshold of at least 20% coverage. We called these gene A and gene B. Third, we filtered for gene pairs encoding proteins that participate in the same complex. As exchangeable subunits of the same complex are likely to be involved in the same biological function, this filter further supported our search for functional compensators. Fourth, we tested the remaining gene pairs for essentiality dependency. We integrated essentiality, expression, copy number and mutation data from cancer cell lines (Methods), resulting in a final dataset that included 878 cell lines with all four types of information. We only retained gene pairs if gene B essentiality was significantly more likely in cell lines with lower expression or damage of gene A (or vice versa). Finally, we ranked the gene pairs based on the number of TCGA samples with loss-of-function alterations. TCGA sample annotations were updated by my colleague Hrvoje Misetic, resulting in a total of 7,953 samples, including 7,921 samples damaged by alterations with a more restrictive definition (Table 7-2, Methods). We expanded the definition of loss of function to include damaging mutations (single nucleotide variants, truncations and indels) in addition to homozygous deletions and double hits.

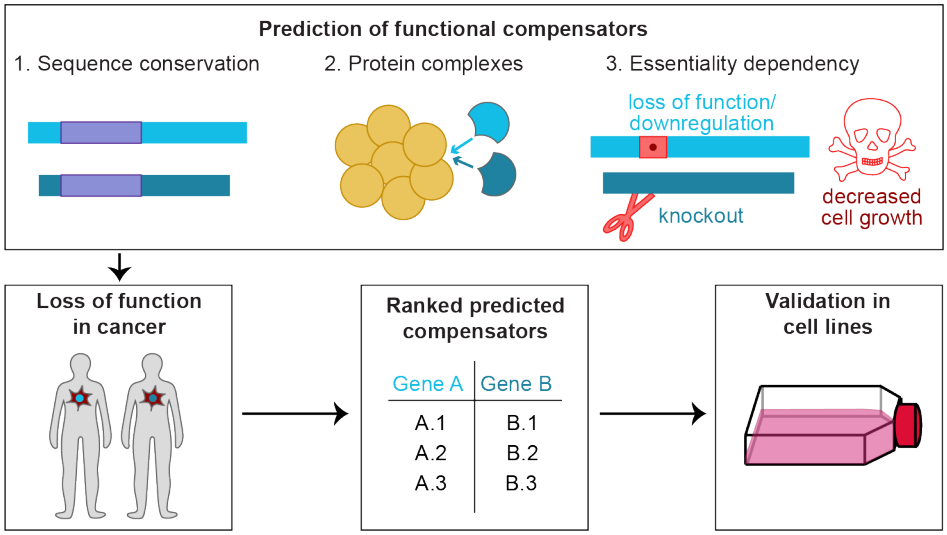


Figure 5-16 Improved workflow for the prediction and validation of functional compensators in cancer genomes.

Potential functional compensator pairs were identified using sequence conservation. Pairs which participated in the same complex were prioritized. In addition, only pairs in which essentiality of one gene was significantly more likely upon loss of function or downregulation of the other gene were chosen. Finally, those genes with recurrent loss-of-function alterations in TCGA samples were prioritized for validation of synthetic lethal interactions in cancer cell lines.

5.6 Improved prediction of functional compensators in cancer genomes

From a total of 19,756 human genes, we identified 1,047 pairs consisting of 2,094 unique genes with at least 20% sequence conservation between both genes (Methods). Out of these, 147 pairs encoded proteins that participated in the same complex and contained at least one context dependent essential gene (i.e., essential in at least one, but not more than 80% of tested cell lines).

We assessed whether essentiality of gene A depended on expression or alteration status of gene B using essentiality, expression, copy number alteration and mutation data from cell lines (Methods). For 52/147 gene pairs, cell lines in which gene A was essential expressed gene B at significantly lower levels (one-sided Wilcoxon test, FDR <0.1). This was the case for known synthetic lethal

pairs *STAG1*/*STAG2* (Figure 5-17A), *SMARCA2*/*SMARCA4* (Figure 5-17B), and the newly identified pair *TBL1X*/*TBL1XR1* (Figure 5-17C), but not for known synthetic lethal pairs *ARID1A*/*ARID1B* (Figure 5-17D) and *CREBBP*/*EP300* (Figure 5-17E). For nine of the 52 gene pairs, the observed dependency was mutual. We identified six gene pairs for which gene A essentiality was significantly more likely in cell lines with a damaging alteration in gene B (i.e., damaging mutation or homozygous deletion, one-sided Fisher test, FDR <0.1). Four of these gene pairs were already identified by the expression level analysis. The two additional pairs were the known synthetic lethal partners *ARID1A*/*ARID1B* and *CREBBP*/*EP300*.

In total, this analysis resulted in 54 gene pairs, which we ranked by their prevalence of loss-of-function alterations in cancer samples from TCGA (Table 7-9). Encouragingly, four of the top five gene pairs were known synthetic lethal pairs *ARID1A*/*ARID1B*, *CREBBP*/*EP300*, *SMARCA2*/*SMARCA4* and *STAG1*/*STAG2* (Table 7-9). Two known synthetic lethal pairs, *EZH1*/*EZH2* and *KMT2A*/*KMT2B,* were not included in the predictions. The annotation of complexes did not contain the participation of EZH1 in complexes, and the sequence conservation between *KMT2A* and *KMT2B* was lower than 20%. *TBL1X*/*TBL1XR1*, the previously validated context dependent synthetic lethal pair, was on the 15th place of the list. *MLLT1*/*MLLT3,* the previously identified false positive synthetic lethal pair, was not included because essentiality was independent of expression and damaging alterations.

Strikingly, the top six predictions were involved in epigenetics. Therefore, we focused on identifying the extent to which epigenetic modifiers were enriched in the predictions, and which evolutionary properties formed the basis of this enrichment. Two of the top six candidates, *MED13*/*MED13L* (Figure 5-17F) and *GATA2*/*GATA3* (Figure 5-17G), were not identified as synthetic lethal pairs previously, and we investigated their synthetic lethal interaction in cell lines.

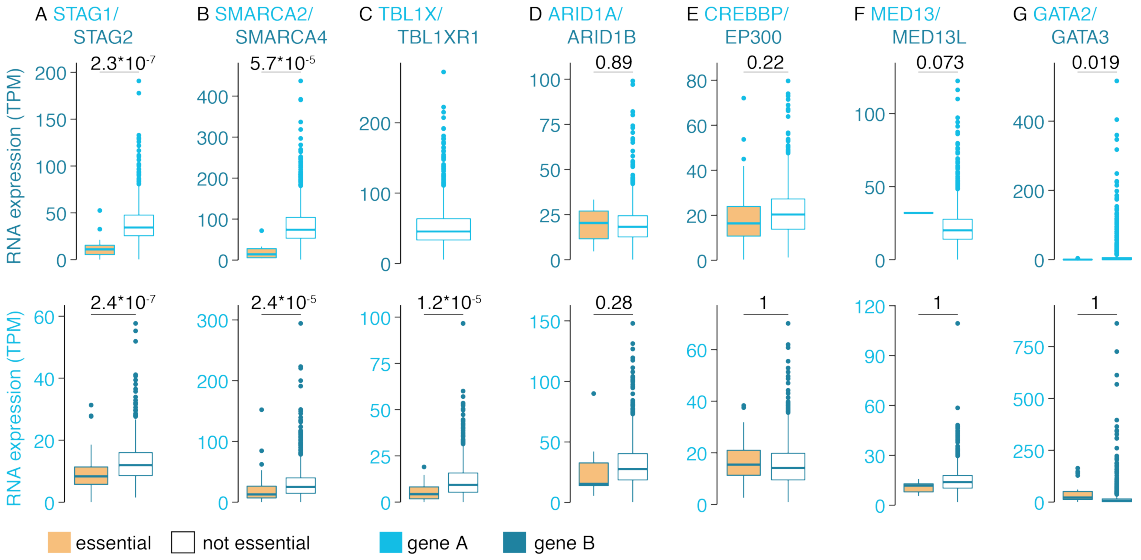


Figure 5-17 Essentiality dependence between gene A and gene B.

For each gene pair gene A/ gene B, only cancer cell lines with available essentiality and RNA expression information were considered. RNA expression of gene B was compared between cell lines in which gene A was essential versus nonessential (top row) and RNA expression of gene A was compared between cell lines in which gene B was essential versus nonessential (bottom row). Gene pairs shown are **A)** *STAG1*/*STAG2*, **B)** *SMARCA2*/*SMARCA4*, **C)** *TBL1X*/*TBL1XR1*, **D)** *ARID1A*/*ARID1B*, **E)** *CREBBP*/*EP300*, **F)** *MED13*/*MED13L*, **G)** *GATA2*/*GATA3*. P-values were calculated using a one-sided Wilcoxon test and adjusted for multiple testing.

To identify the extent and possible reasons for the enrichment of epigenetic modifiers in predicted functional compensator pairs, we compared proportions of 881 epigenetic modifiers retained at each step of the pipeline to proportions of the rest of human genes (18,875 genes). We also compared them to 10,012 genes with functional pathway annotations in Reactome to avoid a bias based on thorough functional characterization. This revealed that a significantly higher proportion of epigenetic modifiers was included in 2,094 unique genes with exactly one protein-coding duplicate with at least 20% conservation (Figure 5-18A). A higher proportion of epigenetic modifiers was also included in the list of 8,371 genes encoding protein complex subunits (Figure 5-18B). Epigenetic modifiers were also more likely to be included in the list of 5,175 context

dependent essential genes (Figure 5-18C), a trend we already observed for chromatin modifiers in Chapter 4.4. We confirmed frequent damage of epigenetic modifiers in cancer samples (Figure 5-18D). Finally, epigenetic modifiers were more likely to be part of the 54 potential functional compensator pairs (Figure 5-18E).

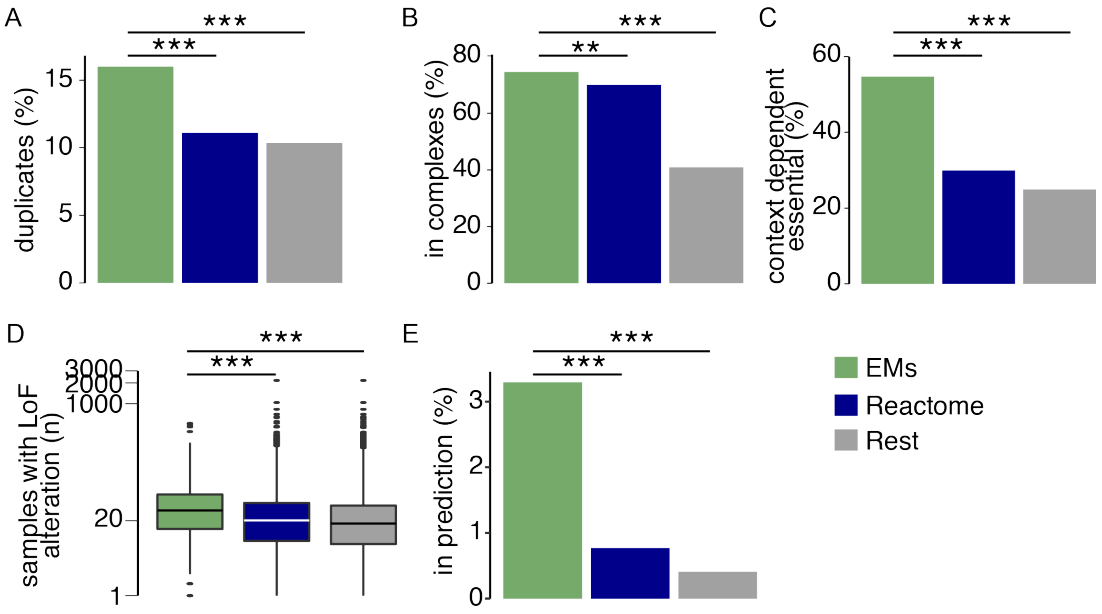


Figure 5-18 Enrichment of epigenetic modifiers at each prediction step.
At each step, the proportion of epigenetic modifiers (total: 881 genes) compared to genes with pathway annotations in Reactome (total: 10,012 genes) and the rest of human genes (non-EMs, total: 18,875 genes) is shown. **A)** Percentage of genes in duplicate pairs including exactly two protein-coding genes with at least 20% sequence conservation. **B)** Percentage of genes participating in complexes. **C)** Percentage of genes which are context dependent essential (essential in at least one cell line, and less than 80% of investigated cell lines). **D)** Distribution of the number of samples in TCGA with a loss-of-function (LoF) alteration. **E)** Percentage of genes in the final list of predicted functional compensators. P-values were calculated using a two-sided Fisher test (A, B, C, E) or Wilcoxon test (D). **p<0.01; ***p<0.001

5.7 Validation of improved predictions of functional compensators in cancer genomes

5.7.1 Validation of synthetic lethality between *MED13*/*MED13L*

MED13 and *MED13L* were damaged in 110 and 92 out of 7,953 TCGA samples, respectively. Based on this high alteration frequency, we chose this gene pair for further validation. Approximately 25% of the protein sequence is conserved between MED13 and MED13L and they are interchangeable subunits of the human mediator complex (Daniels et al., 2013; Sato et al., 2004). Deletions of ten of the 25 mediator complex subunits are lethal in yeast, whereas MED13/MED13L are a part of its regulatory unit containing several interchangeable subunits (Soutourina, 2018). Double knockdown of *MED13* and *MED13L* in colorectal cancer cell lines were shown to decrease cell proliferation (Kuuluvainen et al., 2018); however, the effect of a single gene knockdown was not tested in this context.

MED13 and *MED13L* were damaged in 30 (8.0%) and 23 (6.1%) of 373 endometrial cancer samples, respectively. Due to this high alteration frequency, we chose the two endometrial cancer cell lines HEC1A and TEN to validate this gene pair. Both *MED13* and *MED13L* were expressed in both cell lines (Figure 5-19).

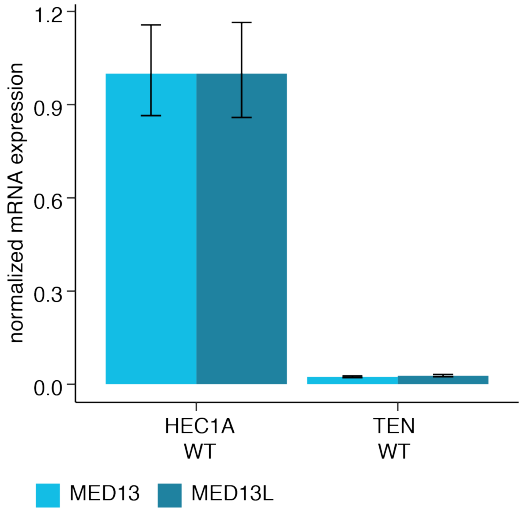


Figure 5-19 Expression levels of *MED13* and *MED13L* in HEC1A and TEN.
RNA expression of *MED13* and *MED13L* was quantified using qPCR in one biological replicate and technical triplicates. Expression levels were normalized to *GAPDH* expression levels in each cell line and normalized to expression levels of the respective gene in HEC1A cells using the ΔΔCt method.

HEC1A and TEN were WT for both genes as determined using alteration and copy number data from DepMap (Chapter 2.2.3). The knockout efficiency of both genes remained above 80% for both cell lines for 14 days (Figure 5-20). We did not observe any difference in proliferation between control, single knockout of *MED13* and *MED13L,* and double gene knockout (Figure 5-21). We confirmed this observation in triplicates at day seven for both cell lines (Figure 5-21A-D), and at day 14 for HEC1A in triplicates (Figure 5-21E,F) and TEN for one replicate (Figure 5-21G,H). We concluded that this gene pair does not engage in synthetic lethal interaction in these two cell lines.

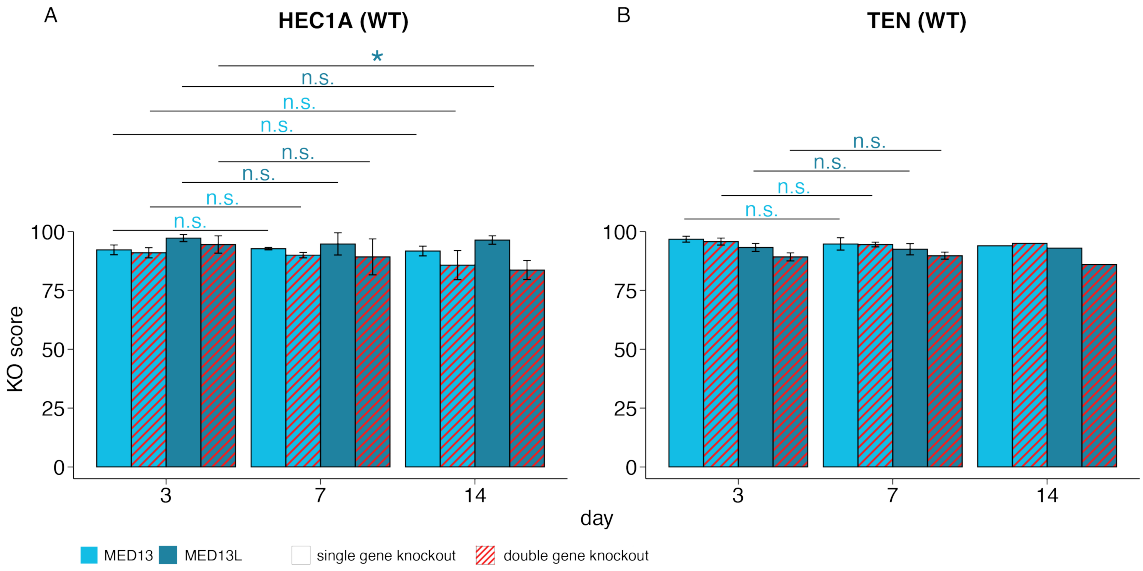


Figure 5-20 Knockout efficiency of *MED13*, *MED13L* or double gene knockout. Recombinant Cas9 protein and sgRNAs were introduced into **A)** HEC1A and **B)** TEN cells using nucleofection. At several time points after nucleofection, DNA was extracted from cells. Loci with expected CRISPR Cas9 editing were amplified by PCR and sequenced using Sanger sequencing. The Synthego ICE tool (Hsiau *et al.*, 2018) was used to assess the percentage of alleles of each respective gene that was altered by a damaging alteration following CRISPR Cas9 editing. KO scores between different time points were compared and p-values were calculated using a two-sided t-test. *p<0.05; n.s. not significant.

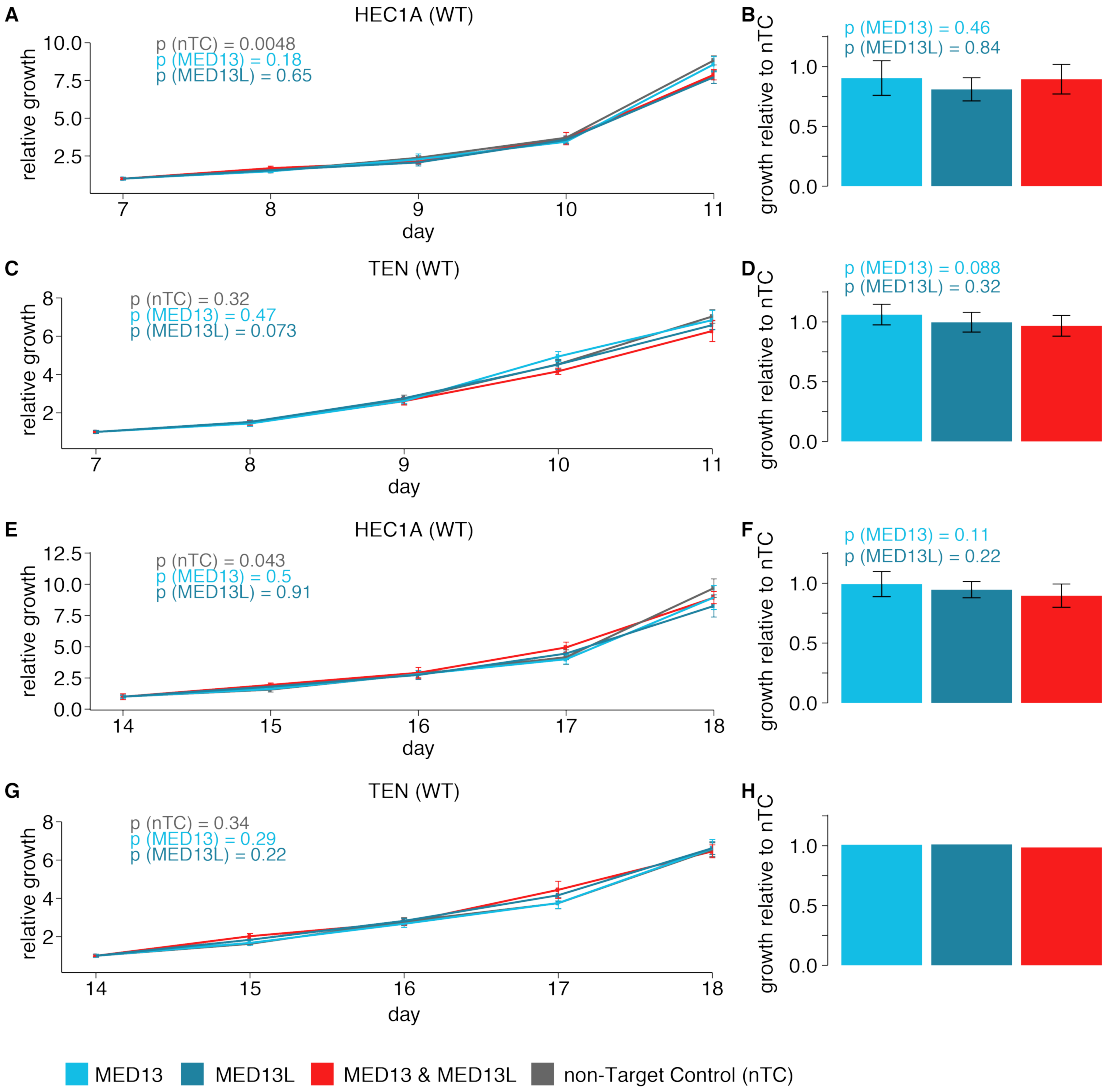


Figure 5-21 Proliferation of cell lines with *MED13*, *MED13L* or double gene knockout.

A,C,E,G) Recombinant Cas9 protein and sgRNAs were introduced into A,E) HEC1A and C,G) TEN cells using nucleofection. After 7 or 14 days, proliferation assays were started and run for four days. A representative proliferation assay out of A,C,E) three replicates and G) one replicate (Figure 7-5) is shown. P values were calculated using a one-sided t test and indicate significance at the last day of the assay between nTC, *MED13* or *MED13L* knockout and the double knockout. **B,D,F,H)** Relative quantification of cell proliferation compared to nTC knockout at the end point. Values represent the mean of three replicates (B,D,F) or the value of a single experiment (H), error bars indicate +/- standard deviation intervals. P values were calculated using a one-sided t test.

5.7.2 Validation of synthetic lethality between *GATA2*/*GATA3*

The second gene pair we chose for validation was *GATA2*/G*ATA3*, with damaging alterations in 38 and 131 out of 7,953 TCGA cancer samples for *GATA2* and *GATA3*, respectively. The GATA transcription factor family comprises six members in humans with two conserved zinc finger domains and a nuclear localization signal (Tremblay et al., 2018). *GATA2* and *GATA3* are each other's most conserved duplicates at approximately 30% sequence conservation. GATA2 and GATA3 recruit the histone acetyl transferase EP300 to target locations (Wu et al., 2014; Yamashita et al., 2002), and GATA3 participates in the Myb complex, a histone acetylation and methylation complex (Nakata et al., 2010). Double gene knockout leads to abnormal prostate development, while single gene knockout does not (Xiao et al., 2016).

GATA3 alterations occurred most frequently in breast cancer samples, with damaging alterations in 72 out of 738 samples (9.8%). Therefore, we chose two breast cancer cell lines, HCC1937 and JIMT1, to validate synthetic lethal interaction between these two genes. Both cell lines were WT for both genes as determined using alteration and copy number data from DepMap (Chapter 2.2.3). *GATA2* and *GATA3* were expressed in both cell lines (Figure 5-22).

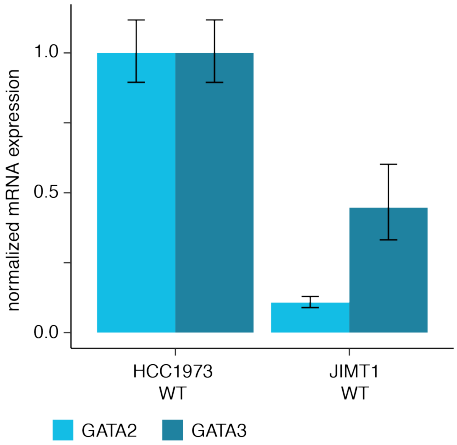


Figure 5-22 Expression levels of *GATA2* and *GATA3* in HCC1973 and JIMT1.
RNA expression of *GATA2* and *GATA3* was quantified using qPCR in one biological replicate and technical triplicates. Expression levels were normalized to GAPDH expression levels in each cell line and normalized to expression levels of the respective gene in HCC1937 cells using the ΔΔCt method.

We observed growth differences as early as seven days after knockout of *GATA2* and *GATA3*. Therefore, compared to previously tested gene pairs, the cell lines were grown in culture for only seven days between the knockout of *GATA2*/*GATA3* and the proliferation assay. In addition, we tested two different experimental approaches. First, we performed a sequential gene knockout. *GATA2* was knocked out first, *GATA3* was knocked out after seven days and proliferation was tested after a further seven days. Second, we knocked out both *GATA2* and *GATA3* in the simultaneous knockout approach and assessed the proliferation rate after seven days. Nucleofections resulted in approximately 70% knockout efficiency in HCC1937 (Figure 5-23A,C) and JIMT1 (Figure 5-23B,D) except for *GATA2* knockout which decreased over time (Figure 5-23A,B,D).

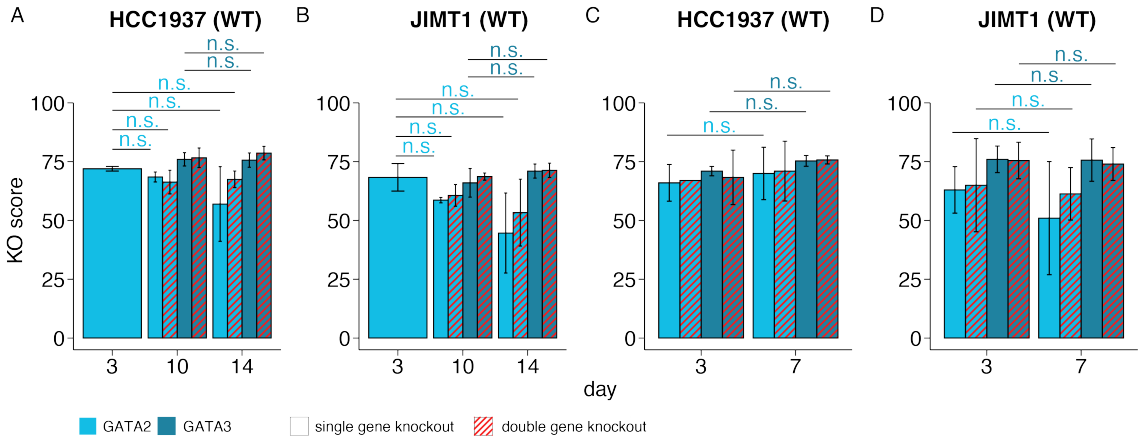


Figure 5-23 Knockout efficiency of *GATA2*, *GATA3* or double gene knockout. Recombinant Cas9 protein and sgRNAs were introduced into **A,C)** HCC1937 and **B,D)** JIMT1 cells using nucleofection. At several time points after nucleofection, DNA was extracted from cells. Loci with expected CRISPR Cas9 editing were amplified by PCR and sequenced using Sanger sequencing. The Synthego ICE tool (Hsiau *et al.*, 2018) was used to assess the percentage of alleles of each respective gene that was altered by a damaging alteration following CRISPR Cas9 editing. KO scores between different time points were compared and p-values were calculated using a two-sided t-test. n.s. not significant.

In all replicates except one, double knockout of *GATA2* and *GATA3* led to reduced proliferation compared to control cells or single gene knockout (Figure 5-24A,C,E,G, Figure 7-6). Considering the average across three replicates at the experimental end point, *GATA2* and *GATA3* double knockout led to significantly reduced proliferation ($p<0.1$) compared to control or single gene knockout in HCC1937 cells (Figure 5-24B,F). We also confirmed this effect in JIMT1 cells as a trend (Figure 5-24D,H), but the comparison between *GATA2* single gene knockout and *GATA2*/*GATA3* double gene knockout was not significant, possibly due to high variation across the three replicates.

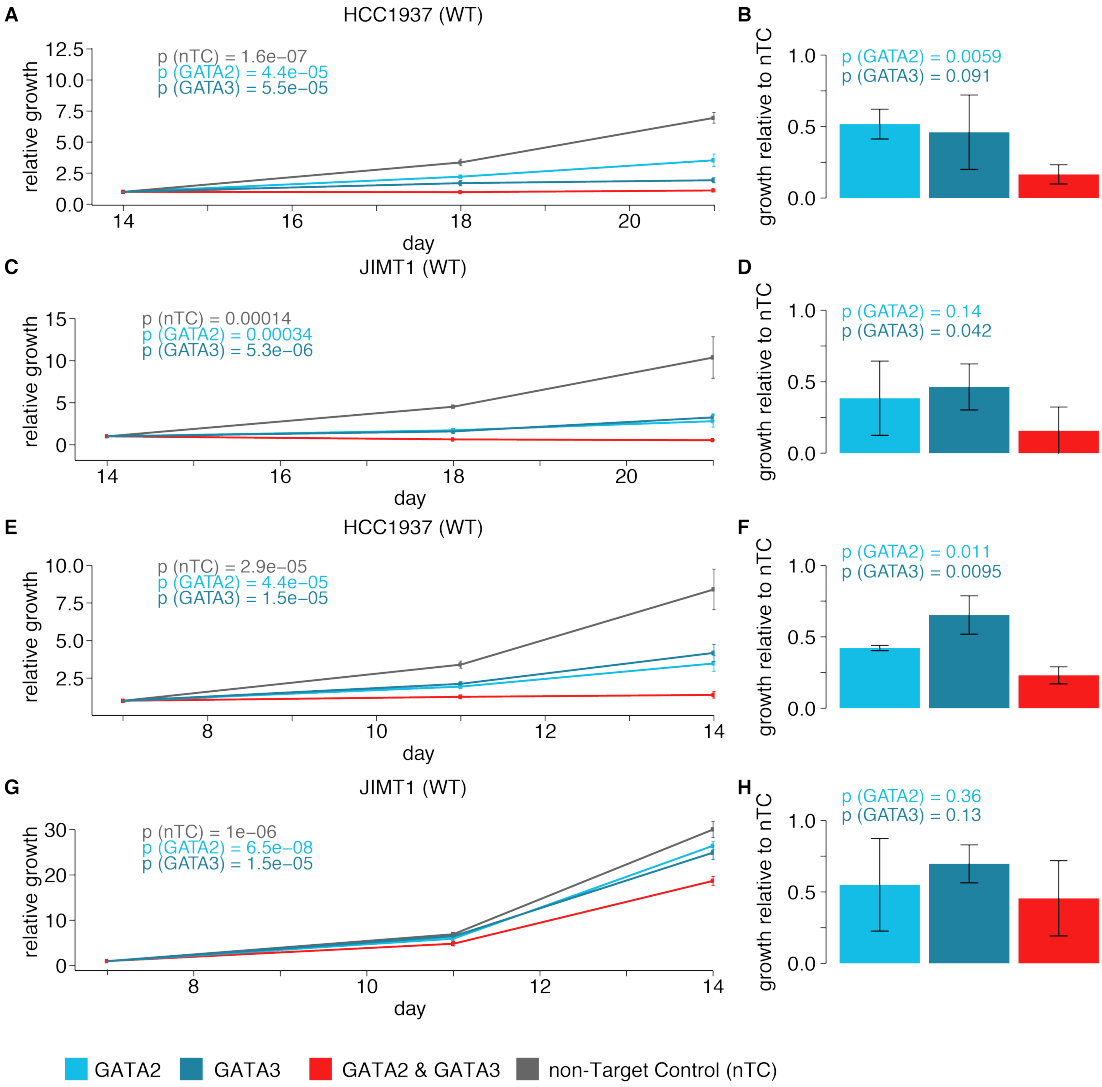


Figure 5-24 Proliferation of cell lines with *GATA2*, *GATA3* or double gene knockout. **A,C,E,G)** Recombinant Cas9 protein and sgRNAs were introduced into A,E) HCC1937 and C,G) JIMT1 cells using nucleofection. *GATA2* was knocked out seven days prior to *GATA3* knockout in A) and C), *GATA2* and *GATA3* knockout was performed simultaneously in E) and G). Seven days after the last nucleofection, proliferation assays were started and run for seven days. A representative proliferation assay out of three replicates (Figure 7-6) is shown. P values were calculated using a one-sided t test and indicate significance at the last assay day between nTC, *GATA2* or *GATA3* knockout and the double knockout. **B,D,F,H)** Relative quantification of cell proliferation compared to nTC cells at the end point. Values represent the mean of triplicates, error bars indicate +/- standard deviation intervals. P values were calculated using a one-sided t test.

5.7.3 Conclusion

Compared to the first method, we expanded the prediction of synthetic lethal pairs by four criteria. First, we predicted synthetic lethal gene pairs across all human protein coding genes. Second, we used a more stringent sequence conservation threshold of at least 20% conservation between both genes. Third, we only included pairs encoding proteins that participate in the same complex. Fourth, we tested whether the essentiality of one gene depended on expression or alteration of its partner.

This improved prediction method identified potential synthetic lethal interactions across all human genes, enabling a comparison between epigenetic modifiers and the rest of human genes. Epigenetic modifiers had characteristic evolutionary properties favouring functional compensation and making epigenetic modifiers a highly interesting target for synthetic lethality-based cancer treatments. These properties included an enrichment in genes with exactly one protein-coding duplicate, genes encoding proteins that participate in a complex, context dependent essential genes and genes with loss-of-function alterations in cancer. The known functional compensators *EZH1*/*EZH2* and *KMT2A*/*KMT2B* were not identified for different reasons. The participation of EZH1 in protein complexes was not annotated in our datasets and we therefore excluded the gene pair. The sequence conservation between *KMT2A* and *KMT2B* was only 14% and the gene pair did not meet the 20% cut-off.

The new method ranked established synthetic lethal epigenetic modifier pairs at the top of the prediction list. In addition, it re-identified the previously confirmed context dependent functional compensator pair *TBL1X*/*TBL1XR1*, while excluding the false positive pair *MLLT1*/*MLLT3*. We identified the synthetic lethal pair *GATA2*/*GATA3*, and a false positive pair, *MED13*/*MED13L*. Stricter p-value cut-offs for the identification of dependency between essentiality of gene A and expression of gene B might help to avoid the identification of false positive gene pairs in the future. The adjusted p-value of true positive functional compensators such as *SMARCA2*/*SMARCA4*, *STAG1*/*STAG2* and *GATA2*/*GATA3* was below 0.02, while it was 0.073 for the false positive pair *MED13*/*MED13L* (Table 7-9). The functional compensation of *MED13* and *MED13L* may also be context

dependent, as a double knockdown decreased proliferation of colorectal cancer cells (Kuuluvainen *et al.*, 2018). Context dependency likely remains an important issue to address for clinical applications. Even for established synthetic lethal pairs, low expression of one partner and knockout of the other partner was tolerated in some cell lines (Figure 5-17A,B,D,E).

Chapter 6. Discussion

6.1 Summary

The work presented in this thesis focuses on the identification and characterization of cancer drivers, and their investigation as potential acquired and cancer-specific vulnerabilities that can be targeted in therapy. As described in the first results chapter, a collection of cancer and healthy driver genes as well as annotations of their evolutionary properties lays the foundation for subsequent analyses. Cancer drivers and healthy drivers have characteristic properties that distinguish them from the rest of human genes, an observation that is followed up in a broader context as described in the second results chapter. The integration of nine evolutionary properties reveals a connection between properties, gene function in healthy cells and the impact of loss-of-function alterations in health and disease. In particular, the integration enables us to prioritize a subgroup of candidate cancer drivers identified from cancer sequencing screens for validation. It also identifies that genes involved in chromatin organization may be enriched in functional compensator pairs that could be exploited in cancer therapy based on synthetic lethality. In the third results chapter, we use a subset of evolutionary properties analysed in Chapter 3 and Chapter 4 to predict synthetic lethal interactions between paralog gene pairs. We confirm the predisposition of epigenetic modifiers to be involved in synthetic lethal interactions. Experimental validation of predictions confirms synthetic lethality between two predicted gene pairs and points towards potential improvements of the prediction workflow.

6.2 The Network of Cancer Genes collects and characterizes drivers in cancer and healthy tissues

The first results chapter describes the update and advancement of the NCG database. Compared to similar published resources (Ainscough *et al.*, 2016; Cerami *et al.*, 2012; Chakravarty *et al.*, 2017; Liu *et al.*, 2020; Tamborero *et al.*,

2018; Tate *et al.*, 2019), NCG focuses on driver genes rather than driver alterations. Unlike other driver gene databases focused on specific groups of drivers (Agarwal *et al.*, 2016; Futreal *et al.*, 2004; Liu *et al.*, 2017; Zhao *et al.*, 2016), NCG comprehensively collects cancer drivers with diverse functions from all tissue types and combines both canonical cancer drivers and candidates identified from cancer sequencing screens. Finally, in contrast to databases purely based on literature-mining (Lever *et al.*, 2019), NCG provides a well curated, manually annotated list of cancer drivers. These unique qualities make NCG a valuable resource for cancer researchers. Frequent updates of the database, two of which were described in this thesis, ensure the representation of our current knowledge on cancer drivers. The most recent update collects 3,177 drivers, a number we expect to increase in the future. As shown by the positive correlation between the number of cancer samples sequenced and cancer drivers identified, driver identification is still limited by the number of sequenced samples. Therefore, additional sequencing of cancer genomes will be needed to identify additional cancer drivers, especially those with low alteration frequency.

In addition to providing a curated list of cancer genes, NCG identifies their evolutionary properties, which are characteristic properties acquired throughout a gene's or protein's evolutionary history. We first investigate previously identified properties of cancer genes. In line with previous results (An *et al.*, 2016; D'Antonio and Ciccarelli, 2011; Domazet-Loso and Tautz, 2008; 2010; Rambaldi *et al.*, 2008), we confirm unique evolutionary properties of cancer drivers, such as their low duplicability in the human genome, early evolutionary origin, frequent interactions with miRNAs, ubiquitous expression, central, highly connected and interconnected position in the PPIN and frequent complex engagement of the encoded proteins. We further expand the set of evolutionary properties by including recent datasets that encompass information on a large proportion of human genes. Only those datasets in which cancer genes differ significantly from the rest of human genes are included in the evolutionary properties of cancer genes. We discover that cancer drivers are more essential in cancer cell lines and less frequently damaged in the germline compared to the rest of human

genes. These properties are independent of the cancer type in which cancer drivers were observed. In the context of the topology of the human PPIN, these properties point towards cancer genes as vulnerable hubs, whose loss leads to malfunction of the network. Candidate cancer drivers have similar but less pronounced properties compared to canonical drivers. Among all candidates, those supported by at least two sequencing screens exhibit stronger similarities to canonical cancer drivers, indicating that false positives may exist among the remaining candidates. This observation motivated us to identify likely true positive candidates using an integrated analysis of evolutionary properties, as described in the second results chapter. The characteristic evolutionary properties of cancer drivers can also be used to identify novel cancer drivers. Applying a machine learning approach to cancer sequencing data, Mourikis *et al.* identified patient-specific drivers with properties similar to canonical cancer gene properties (Mourikis *et al.*, 2019; Nulsen *et al.*, 2021).

In the latest version of NCG, we annotate driver genes identified through alterations in their noncoding regions, and the drivers of clonal expansion in non-cancer tissues. Cancer drivers with noncoding alterations do not have properties as pronounced as drivers with coding alterations. This indicates a less essential role in the cell of drivers with noncoding alterations, as supported by their higher tolerance towards germline variation and less central position in the PPIN. These drivers possibly contribute to the progression of cancer through a different mechanism. However, only 721 of 3,177 drivers identified through sequencing screens are altered in noncoding regions. Thus, these results need further investigation once more drivers with noncoding alterations are identified, for example by performing additional whole genome cancer sequencing screens. The development of new methods dedicated to the detection of driver alterations within noncoding regions of the human genome will support this effort.

Drivers identified in both cancer and non-cancer tissue exhibit the strongest difference from the rest of human genes, while the remaining healthy drivers have different evolutionary properties compared to cancer drivers. Considering the PPIN, these properties indicate a central, vulnerable hub position for healthy cancer drivers and a peripheral position with higher tolerance towards mutations

for the remaining healthy drivers. This points towards two different groups of genes driving clonal expansion. Healthy drivers which have never been observed in the context of cancer may have different evolutionary properties because they can exclusively drive clonal expansion in non-cancer tissues. Alterations of canonical healthy drivers may initially drive the expansion of non-cancer clones which evolve into malignant tumours under the prerequisite of additional factors. For example, accumulation of a sufficient number of altered canonical healthy drivers over time may be necessary (Stratton *et al.*, 2009). A further requirement for cancerous transformation may be the correct order of co-occurring alterations (Lee-Six *et al.*, 2019) and additional copy number alterations and chromosomal rearrangements. These additional driver categories may provide useful additional information if they are added to the NCG resource in the future. Conclusions regarding healthy drivers need to be confirmed once further healthy drivers are identified, as the group of remaining healthy drivers was only comprised of eight genes.

Another interesting hint deriving from these studies is that some genes included in the canonical healthy driver category may not actually contribute to cancer progression. Current driver identification methods are based on a higher alteration frequency in cancer than expected and do not account for positive selection before cancer initiation. Thus, it is possible that some alterations accumulate due to positive selection in healthy or pre-malignant tissue and are therefore overrepresented in the cancer originating from this tissue, without being involved in cancer initiation or progression. Continuous improvement of cancer driver identification methods is necessary and experimental validation of identified driver genes remains of high importance. Multiplex assays interrogating the effect of multiple DNA alterations on biological function are accelerating the validation of alterations linked to disease (Findlay, 2021). For example, deep mutational scanning introduces a library of mutated gene variants into cells where the function of the encoded protein provides a selective advantage (Fowler and Fields, 2014). DNA sequencing is then used to identify the variants that result in functional protein. Saturation editing of genomic regions using CRISPR Cas9 with a library of donor templates for homology-directed repair enables the assessment

of alterations in the original context of the DNA locus (Findlay et al., 2014). The consequence of non-coding alterations can be identified by parallel editing of enhancers and subsequent RNA sequencing (Melnikov et al., 2012; Patwardhan et al., 2012).

In summary, the identification of drivers with noncoding alterations and healthy drivers is a promising field in cancer research. Increasing the number of cancer sequencing screens will be essential in gaining more power to identify cancer drivers, especially rare drivers. In this context, the focus should shift to whole genome sequencing screens to enable the identification of drivers with noncoding alterations. The identification of cancer drivers needs to be supported by novel driver identification methods tailored to noncoding alterations and factoring in selection in the healthy tissue. Experimental validation remains crucial to understand the biological role of drivers.

In the future, NCG will continue to grow its collection of drivers. Especially for drivers with noncoding alterations and healthy drivers, the increase in genes will help to refine the picture of their evolutionary properties. It will also be interesting to add further properties to the repertoire to gain a more complete picture. These datasets need to be comprehensive to enable a thorough comparison between the different categories of cancer genes and the rest of human genes. For example, post-translational modifications may be included as evolutionary properties once the high-confidence data on post-translational modifications such as phosphorylation is available for the majority of human genes. It may also be interesting to investigate properties that are characteristic of other gene groups. For example, human essential genes are characterised by a certain gene length and chromatin accessibility (Chen *et al.*, 2020). Finally, new, more precise annotations of evolutionary properties, for example protein complexes or miRNA interactions, will further improve our characterization of driver genes.

6.3 Integrating evolutionary properties identifies functional gene groups and prioritizes cancer driver candidates

Inspired by the distinctive property profile of cancer driver genes, the second results chapter focuses on identifying characteristic properties of gene groups in a wider context. To enable a comprehensive comparison, we develop the EP score and show that it captures the majority of differences between 25 functional pathways regarding nine evolutionary properties. High EP scores indicate a combination of properties pointing towards a central role of genes in the cell. Accordingly, the EP score reveals similar evolutionary properties for genes within the same pathway and groups pathways according to function: Pathways with high EP scores are involved in basic cellular functions, whereas pathways with low EP scores are involved in specialized functions.

Essential genes have high EP scores and are, as expected, enriched in pathways with low tolerance to germline and somatic alterations. In concordance with individual property differences shown in Chapter 3.2.2 and Chapter 3.3.2, cancer drivers also have high EP scores. Interestingly, the highest EP scores are observed for cancer genes which are also healthy drivers, indicating their involvement in basic cellular functions. Given their high EP scores, it is not surprising that essential genes and cancer drivers are overrepresented in functional groups with high EP scores. Interestingly, chromatin organization and circadian clock have the highest proportion of cancer drivers and the lowest percentage of core essential genes among the high-scoring pathways. In contrast, while context dependent essential genes are depleted in the circadian clock pathway as expected, they are enriched in the chromatin organization pathway. This enrichment might be caused by synthetic lethal interactions between functional compensators of the chromatin organization pathway. In addition, this pathway has a high proportion of tumour suppressors, a gene group which is difficult to target. Synthetic lethal interactors within this gene group may represent potential new targets for cancer therapy. We follow up on this hypothesis in Chapter 5.

In addition to cancer drivers, Mendelian disease genes constitute a second group of genetic disease genes. Their EP score is lower compared to cancer drivers, pointing towards a less central role in the cell. Alterations in the germline need to be tolerable while affecting every cell of the developed organism. At the same time, there is no proliferative competition between neighbouring cells regarding a germline alteration, as they all carry the same alteration. In contrast, somatic alterations occur in a single cell of a developed organism. The cell itself is dispensable to the proper function of the organism. It is therefore less restricted in the acquisition of alterations, but to transform into a cancerous clone, it needs to have an advantage over its wild type neighbours. This can be achieved through alterations in central cellular pathways with high EP scores. Dominant Mendelian disease genes have higher EP scores than recessive genes. In concordance, disease genes whose loss is embryonic lethal in mice are more connected in the PPIN (Goh *et al.*, 2007). As a homozygous alteration in the germline of these genes is lethal, they must elicit a phenotype through heterozygous alterations, meaning they have to be dominant disease genes. Given their significantly higher EP score, we did not expect the negative correlation between the percentage of Mendelian disease genes and the pathway EP scores. It indicates that Mendelian diseases tend to be caused by alterations of relatively highly connected nodes within the peripheral gene groups of the PPIN, highlighting a fundamental difference between germline and somatic genetic disease.

The integrated analysis of nine properties with a PCA enables a more granular comparison of genes within a pathway. While all other pathways are composed of genes with similar properties, the reproduction pathway consists of two distinct groups. This is due to a functional difference between the two groups, with one group of genes involved in cell cycle regulation during meiosis, and the other in spermatogenesis. This demonstrates a successful separation of different functional groups of genes within pathways. Similarly, we observe two distinct groups within the candidate cancer driver genes. They are characterized by differing EP scores and clearly separate within the PCA plot. One of these groups overlaps in the PCA with the canonical cancer genes, indicating that these potential cancer drivers should be prioritized for experimental validation. In line

with their different individual properties, genes promoting the clonal expansion of normal cells without leading to cancer transformation have low EP scores, indicating a fundamental evolutionary difference between genes promoting clonal expansion in health and disease. Therefore, the EP score may also support the distinction between healthy and cancer drivers.

This study is limited by the annotation of evolutionary properties and function. We observe similar results using functional annotations by the KEGG database, confirming that we observe a true biological signal. Only approximately half of human genes are annotated in the Reactome functional database (Fabregat *et al.*, 2018), and annotations are possibly biased in favour of well-known genes (Haynes et al., 2018). Similarly, despite our efforts to integrate several databases to achieve the most complete picture possible, annotations of evolutionary properties are still incomplete. While gene essentiality, RNA expression and evolutionary origin are available for approximately 95% of human genes, interactions of encoded proteins are annotated for approximately 80%, miRNA interactions for 75% and protein expression for 65% of human genes (Methods Chapter 2.1.1). Functional annotation is only available for 58% of genes, highlighting the importance of further investigations of protein function. In the future, additional, high-quality annotations will improve our understanding of gene functions and their evolutionary differences.

6.4 Evolutionary properties of epigenetic modifiers predispose them to paralog synthetic lethality

The annotation of gene evolutionary properties forms the basis for the third part of this thesis, the identification of paralog synthetic lethal pairs. We use gene duplicability, interaction of proteins in complexes and gene essentiality in cell lines to predict synthetic lethal paralog pairs. In addition, we determine clinical relevance using the frequency of gene alterations in TCGA.

Epigenetic modifier genes are frequently altered in cancer (Feinberg *et al.*, 2016) and several of them are involved in known synthetic lethal interactions (Chapter 1.4.3). As discussed in Chapter 4, genes involved in chromatin organization may

be enriched in additional synthetic lethal interactors. Therefore, we investigate in Chapter 5 whether epigenetic modifiers are indeed enriched in synthetic lethal interactors and identify potential reasons for this enrichment. As the definition of epigenetic modifier depends on the source and the context, we integrate four sources of epigenetic modifiers and confirm their involvement in epigenetics with a literature search. This list may not be entirely comprehensive and could be extended in the future. For example, eight genes with diverse functions beyond epigenetics which are currently excluded from the analysis could be included. The list could also be restricted to include only catalytic subunits of complexes involved in epigenetics instead of all subunits.

When using the first approach of predicting synthetic lethal paralog pairs, we only consider epigenetic modifiers with frequent loss-of-function alterations in TCGA (homozygous deletions or double hits of heterozygous deletions and damaging alterations). Among these, we identify genes with exactly one coding duplicate gene with at least 5% sequence conservation. The experimental validation confirms only the context dependent synthetic lethal pair, *TBL1X*/*TBL1XR1*, and identifies one pair which does not engage in synthetic lethality, *MLLT1*/*MLLT3*. This may be partly due to experimental limitations as discussed below, and partly due to limitations of the prediction approach, which we address by substantially modifying the prediction pipeline. First, we increase the sequence conservation threshold to 20%. This is in line with a recent study of paralog synthetic lethality, in which only paralog pairs with amino acid sequence identity of at least 20% were retained (De Kegel *et al.*, 2021). Second, we only retain gene pairs that participate in the same protein complex. This is based on the striking observation that most known synthetic lethality pairs are interchangeable subunits of the same complex. Third, given that *MLLT1*/*MLLT3* is not necessary for the proper function of the *DOT1L* complex, it is not surprising that the loss of both genes does not have a negative effect on cell survival. To identify gene pairs relevant for cell survival, we filter for gene pairs where essentiality of one of the partner genes is dependent on mutation or under-expression of the other partner based on cancer cell line data. Fourth, we apply a less restrictive definition of loss-of-function alterations in TCGA and include not only homozygous deletions and

double hits, but also damaging single nucleotide variants, indels and truncations. We did not include gene silencing through DNA methylation as a means of loss of function, since we considered it less precise than copy number alterations or mutations. Methylation of CpG islands in promotors is known to silence expression (Bird, 1992; Keshet *et al.*, 1986), however, the exact assignment of a promotor to its regulating gene may be imprecise. In addition, the downregulation of expression may not lead to a complete loss of function. Further uncertainty results from the fact that the effect of DNA methylation depends on its location. Methylation of some elements such as promoters, the first intron (Anastasiadi *et al.*, 2018) or retroelements (Schulz *et al.*, 2006) leads to gene silencing, while methylation in other parts of the gene body may result in increased expression (Aran *et al.*, 2011). Once the knowledge on gene silencing through DNA methylation increases, it will be interesting to include it as a means of loss of function.

In summary, this pipeline combines approaches proven to be effective predictors of synthetic lethality with indication of clinical relevance. Both sequence conservation and participation in complexes were shown to be among the top indicators of synthetic lethal interactions (De Kegel *et al.*, 2021). Our approach of requiring the partners to engage in the same complex is even more stringent. Of note, genes that are each other's closest paralog were also shown to be likely synthetic lethal interactors (De Kegel *et al.*, 2021). Our filter of gene pairs including exactly two genes is more stringent. Dependency of the essentiality of one partner on expression or mutation status of the other partner proved to be an effective predictor of synthetic lethality in the DAISY approach (Jerby-Arnon *et al.*, 2014). Filtering for genes with frequent loss-of-function alterations in TCGA samples ensures that we identify candidate pairs with high relevance to cancer therapy.

We apply the prediction to all human genes instead of only epigenetic modifiers. This allows us to identify all interesting new candidates and also enables a comparison between epigenetic modifiers and the rest of human genes. Epigenetic modifiers are highly likely to engage in paralog synthetic lethality due to their evolutionary properties. They are more likely to have exactly one duplicate

at a 20% conservation threshold and they engage more frequently in protein complexes. As already indicated for chromatin modifiers in Chapter 4, we find that epigenetic modifiers are more often context dependent essential, and they are more frequently lost in cancer samples. Overall, we confirm our hypothesis that epigenetic modifiers are enriched in synthetic lethal interactors.

While the modifications to the prediction approach are partly successful and result in the validation of *GATA2*/*GATA3* as a synthetic lethal gene pair, the prediction also identifies a false positive gene pair, *MED13*/*MED13L*. Known synthetic lethal partners such as *SMARCA2*/*SMARCA4* or *STAG1*/*STAG2* have very low p-values regarding conditional essentiality (Table 7-9) and exhibit mutual essentiality dependency. Thus, a more stringent threshold for the p-value of the conditional essentiality may improve predictions. Following this idea, *PPP2R1A*/*PPP2R1B*, *ACO1*/*IREB2* or *ELMO1*/*ELMO2* may be interesting candidates. Based on a further restriction of predictions to only include those gene pairs with mutual essentiality dependency, interesting candidates may be *DDX5*/*DDX17* and *VPS4B*/*VPS4A*. Reassuringly, both gene pairs were recently shown to be synthetic lethal gene pairs in the HAP1 cell line (De Kegel and Ryan, 2019). In the future, it will be interesting to validate synthetic lethality between both epigenetics-related and non-epigenetics-related pairs in cell lines (Table 7-9).

The validation screen of nine potential functional compensator pairs is mainly limited by its short duration. Confluency of cells is reached after three to four days (Figure 5-5), while a synthetic lethal effect in subsequent CRISPR Cas9 knockout experiments is only observed after seven to 14 days. Similarly, the study validating *EZH1* and *EZH2* inhibitors only observed an effect on tumour growth 40 days after treating tumour-bearing rats (Honma *et al.*, 2017). The synthetic lethal effect between these two genes may also be cell line dependent, as its potential to slow down cell line proliferation was observed predominantly in leukaemia, lymphoma and myeloma (Honma *et al.*, 2017). It may also depend on cell culture conditions such as pH or oxygen. This may explain why we did not observe synthetic lethality between the *EZH1*/*EZH2* control pair. Given the failure of the positive control, we cannot conclude the absence of synthetic lethality

between the remaining gene pairs tested in this screen. This may be addressed in the future by splitting the cells during the screen, thus allowing for a longer time between knockout and proliferation assessment.

Here, we address this problem by a single gene pair validation approach. We culture cells for seven to 14 days between the CRISPR Cas9 knockout and assessment of proliferation. We see synthetic dependency between *TBL1X*/*TBL1XR1* in one cell line, and between *GATA2*/*GATA3* in two cell lines. While knockout of *TBL1X* in *TBL1XR1*-deficient cells leads to a consistent, significant decrease in proliferation, the double knockout of *GATA2* and *GATA3* is more challenging. When comparing double- to single-gene knockout in individual experiments, we are able to observe significantly reduced proliferation in all conditions except one replicate of JIMT1 (Figure 7-6B). However, proliferation rates after single gene knockout of *GATA2* and *GATA3* are variable, especially in JIMT1 cells, where we also observe a reduction of knockout efficiency of *GATA2* over time. This leads to variability between proliferation rates after single and double gene knockout. Therefore, we only observe a non-significant trend in JIMT1 cells for the overall decrease in proliferation (Figure 5-24). Consequently, the validation in cell lines with homozygous deletions of one gene is preferable. Stable knockout of one gene could be an alternative to reduce experimental noise. In addition, introducing the Cas9 protein and synthetic gRNAs into the cells using nucleofection may not be sufficiently efficient, and the electroporation may be harmful for certain cell lines. Exploring different experimental approaches, such as introducing the Cas9, gRNAs and a selectable marker via a lentiviral system, may improve knockout efficiencies.

We approach the validation of *GATA2*/*GATA3* with two different options, the simultaneous knockout of both genes and the sequential knockout of one gene at a time. Both lead to comparable results regarding knockout efficiency and reduction in proliferation. To make experiments more time efficient, and to reduce the stress on cells through the second nucleofection and the time in which one gene knockout might be lost, we apply simultaneous double knockout for the remaining double knockout experiments.

In summary, the validation of four gene pairs results in one gene pair with consistent synthetic negative interaction (*GATA2*/*GATA3*), one context dependent synthetic lethal pair (*TBL1X*/*TBL1XR1*) and two gene pairs for which we do not observe negative synthetic interactions (*MLLT1*/*MLLT3*, *MED13*/*MED13L*). *MLLT1*/*MLLT3* are not required for the function of the DOT1L complex and essentiality of one does not depend on expression or mutation status of the other in cancer cell lines. Therefore, this gene pair is not likely to engage in negative synthetic interactions. In contrast to our results, double knockout of *MED13*/*MED13L* leads to reduced proliferation in colorectal cancer cell lines (Kuuluvainen *et al.*, 2018). Therefore, this gene pair is likely to engage in context dependent synthetic negative interactions. Testing additional cell lines and investigating the biological effect of *MED13*/*MED13L* double knockout on the function of the mediator complex in different genetic backgrounds will be interesting to investigate in the future.

To identify whether *TBL1X*/*TBL1XR1* may represent a good therapeutic target, synthetic lethal interactions between the two genes need to be identified in additional cell lines, ideally cell lines with homozygous deletions or stable knockouts of one partner. The choice of these cell lines may be guided by the essentiality dependency analysis (Chapter 5.6). Differences between cell lines that do and do not exhibit synthetic lethal interactions between the two genes can then be identified and guide stratification approaches for potential therapies. For example, differential RNA expression analysis may identify a biological background favouring synthetic lethality. As TBL1X and TBL1XR1 interact with HDAC3, leading to deacetylation of histones, a Western Blot of their target marks such as H3K27ac may identify whether additional compensation pathways exist.

GATA2/*GATA3* are synthetic lethal interactors in both of the tested cell lines but need to be validated in additional cell lines and, given a successful validation, in mouse models. The main concerns regarding a potential clinical application of *GATA2* or *GATA3* inhibitors may be their side effects in healthy cells. Although knockout of both *GATA2* and *GATA3* led to the most drastic reduction in proliferation, single gene knockout also reduced cell growth. Encouragingly, the *GATA2* inhibitor *dilazep* was identified as a therapeutic agent in *GATA2*-

dependent prostate cancer and successfully applied in mouse models (Kaochar et al., 2021).

Regarding a potential future use as therapeutic targets in cancer therapy, the identification of synthetic lethal relationships in cell lines is only the first step of many. In general, 2-dimensional cell cultures are only an imprecise model and often do not represent tumour vulnerabilities (Yu et al., 2019). Cells have adapted to growth in artificial conditions over time and can only approximately model growth in three dimensions. They might also not be representative for additional selective mechanisms *in vivo* such as the influence of the tumour environment, hypoxia, pH or nutrient restriction (Casciari et al., 1992; Strese et al., 2013; Yu *et al.*, 2019). In addition to cell line related challenges, synthetic lethal interactions often depend on the context they occur in (Chapter 1.3.4). Therefore, identified targets need to be thoroughly tested in animal models and clinical trials in humans. Despite the high interest and discovery of many synthetic lethal gene pairs in the past, only the synthetic lethal interaction between PARP inhibitors and *BRCA1* or *BRCA2* mutations has succeeded as a therapeutic approach so far. This demonstrates the effort that is still needed to successfully exploit synthetic lethality for cancer treatment.

While cancer genomics has greatly improved our ability to stratify patient groups, genomics alone cannot fully capture all aspects that may have an impact on therapy response (Letai, 2017). Intra-tumour heterogeneity limits the benefit of targeted therapies, as sub-clones without the alteration conferring vulnerability may be resistant to the therapy (McGranahan and Swanton, 2017). Targeting clonal loss-of-function alterations and a combination of several synthetic-lethality-based therapies could address this issue. High throughput approaches for individual tumours are able to predict response to individual therapies, for example using multiplexed drug testing of tumour slices (Horowitz et al., 2020) or *ex vivo* cultures of tumour derived cells (Meijer et al., 2017). The efficacy of up to eight anti-cancer drugs can also be tested directly in a solid tumour before it is resected from the patient (Gundle et al., 2020; Klinghoffer et al., 2015). These approaches may help to identify patients who would benefit from therapies targeting context dependent synthetic lethal interactors.

6.5 Conclusion

Since the sequencing of the first cancer genome in 2008 (Ley et al., 2008), a global sequencing effort has vastly increased our understanding of cancer. This thesis contributes to our understanding of cancer genes. The integration of their evolutionary properties allows a comprehensive analysis of their role in the cell regarding functional gene groups in health and disease. It also identifies likely true positive candidate cancer genes. Finally, a subset of evolutionary properties enables the prediction of paralog functional compensator pairs. These are enriched in epigenetic modifiers due to their evolutionary properties predisposing them to paralog functional compensation. Among the predicted therapeutic targets, we validate synthetic lethality between two pairs.

In the future, the expansion of sequencing screens in non-cancer tissues will continue to reveal differences and parallels between healthy and cancer drivers. This may be supported by sequencing healthy tissue biopsies from donors who passed away due to non-cancer-related causes. The sequencing of additional cancer genomes, especially whole genome sequencing, is expected to provide exciting new insights. In a clinical setting, the collection of sequencing data as part of personalized treatment is becoming increasingly affordable and accepted. New approaches such as DNA sequencing from liquid biopsies may provide a routine way to track cancer mutations over time, facilitating the identification of personalized therapeutic targets. In this context, targeting functional compensators of lost genes in cancer represents a promising approach. With the identification of more functional compensator pairs, therapies could simultaneously exploit several targets thanks to low toxicity in healthy cells. If clinical data are made available for research, they will present a great resource in addition to datasets from sequencing consortia.

Chapter 7. Appendix

7.1 Supplementary Tables

Table 7-1 Annotation of median EP score, loss-of-function alterations, and involvement in disease for 10,334 member genes of 25 pathways.

Each of the 25 Reactome level 1 functional pathways used in the analysis is listed with its pathway group according to median EP score (see Figure 2B), its number of included genes, the median EP score and median LOEUF score. From all genes involved in the respective pathway, the median number of times a gene is altered by a loss-of-function alteration (homozygous deletion, truncating mutation or damaging mutation) was counted in 7,626 cancer samples from TCGA and 1,234 cell lines from the Cancer Cell Line Encyclopedia (CCLE). The percentage of essential genes, cancer drivers, healthy drivers and Mendelian disease (MD) genes is provided for each pathway. LoF - loss-of-function, TSG – tumour suppressor gene, OG – oncogene

Pathway	Reactome group	genes (n)	EP score	median LOEUF score	median genes (n) with LoF alterations		Essentiality: Percentage (%) genes that are		Cancer genes: Percentage (%) genes that are			Healthy drivers: Percentage (%) genes that are		Mendelian disease genes: Percentage (%) genes that are		
					TCGA	CCLE	Core essential	Context essential	Cancer drivers	TSG	OG	Only healthy	Healthy and cancer	MD gene	Recessive MD gene	Dominant MD gene
Circadian Clock	1	49	0.84	0.34	15	13	0.00	34.04	20.41	6.12	8.16	0.00	4.08	24.49	4.08	14.29
Chromatin organization	1	274	0.82	0.41	13	9	2.62	67.79	16.06	6.20	5.84	0.36	3.28	23.36	4.01	18.25
DNA Repair	1	310	0.82	0.81	9	7	10.49	56.72	14.84	12.26	0.65	0.00	1.61	25.48	17.10	4.52
Cell Cycle	1	636	0.82	0.64	10	7	14.67	56.78	11.79	4.72	3.77	0.47	1.42	23.11	12.42	8.33
Programmed Cell Death	1	170	0.84	0.47	10	7	13.61	35.50	10.59	4.71	2.94	0.00	2.35	22.94	5.29	11.18
Cellular responses to external stimuli	1	501	0.83	0.61	9	6	8.72	48.88	9.38	2.99	3.39	0.40	1.00	17.96	6.19	10.18
DNA Replication	1	107	0.84	0.63	9	7	38.32	47.66	5.61	4.67	0.00	0.00	1.87	15.89	10.28	5.61
Metabolism of RNA	1	662	0.84	0.54	8	6	27.76	56.75	4.98	1.51	1.81	0.30	0.45	18.43	10.42	6.80
Mitophagy	1	29	0.84	0.80	6	5	6.90	41.38	3.45	0.00	3.45	0.00	0.00	20.69	6.90	6.90
Gene expression (Transcription)	2	1384	0.69	0.64	11	7	7.07	43.59	12.36	4.84	4.34	0.29	1.73	19.00	6.36	9.75

Developmental Biology	2	1052	0.66	0.58	12	8	8.43	34.00	9.98	2.57	4.18	0.57	1.43	28.14	5.89	16.63
Hemostasis	2	643	0.64	0.64	13	10	0.62	26.37	8.40	1.09	4.82	0.47	1.56	32.35	11.35	13.37
Immune System	2	1967	0.66	0.76	10	7	3.60	27.65	7.52	1.98	3.25	0.10	0.81	24.56	12.15	7.88
Organelle biogenesis and maintenance	2	294	0.71	0.74	13	10	6.14	38.23	7.14	0.68	2.04	1.36	0.68	38.44	23.81	10.20
Vesicle-mediated transport	2	669	0.69	0.58	12	8	3.63	34.14	5.53	1.20	1.49	0.30	0.15	30.19	13.15	11.66
Metabolism of proteins	2	2005	0.68	0.72	9	7	8.48	39.30	5.19	2.24	1.15	0.15	0.50	25.94	15.36	7.28
Reproduction	3	144	0.52	1.10	8	6	4.38	50.36	12.50	7.64	2.78	0.00	0.69	23.61	12.50	6.94
Cell-Cell communication	3	129	0.52	0.45	19	12	0.00	25.58	11.63	4.65	5.43	0.78	2.33	37.21	13.95	13.95
Signal Transduction	3	2608	0.52	0.72	12	8	2.60	25.19	8.70	2.65	3.80	0.19	1.11	22.70	7.59	10.31
Extracellular matrix organization	3	298	0.51	0.66	22	16	0.00	12.46	4.70	1.01	1.34	1.01	0.00	47.32	19.46	16.11
Muscle contraction	3	205	0.47	0.63	14	11	0.49	14.71	4.39	0.98	1.46	0.00	0.00	46.83	5.85	27.80
Neuronal System	3	367	0.37	0.44	22	14	0.00	12.53	4.09	0.54	2.45	0.00	0.00	35.69	9.26	20.44
Metabolism	3	2048	0.52	0.91	11	7	4.77	29.43	3.22	1.27	0.78	0.15	0.39	34.03	24.80	5.47
Transport of small molecules	3	720	0.48	0.80	15	11	4.59	19.61	1.81	0.56	0.56	0.42	0.00	32.36	17.22	8.61
Digestion and absorption	3	27	0.16	1.19	19	11	0.00	4.17	0.00	0.00	0.00	0.00	0.00	22.22	18.52	0.00

Table 7-2 Overview of samples in TCGA.

Copy number alterations, damaging mutations and RNA expression were updated while working on several projects. The first annotation was obtained from Thanos Mourikis for 7,828 TCGA samples and used for the first prediction of synthetic lethal interactors. The second annotation was obtained from Hrvoje Misetic for 7,921 samples (7,626 samples upon exclusion of hypermutated samples). These were used for NCG7, analysis of evolutionary properties and the second prediction of synthetic lethality.

Cancer type	Annotation by Thanos Mourikis (n)	Annotation by Hrvoje Misetic (n)	Annotation by Hrvoje Misetic, filtered (n)
Adrenocortical Carcinoma	72	77	76
Bladder Urothelial Carcinoma	232	363	346
Breast Invasive Carcinoma	954	738	726
Cervical squamous cell carcinoma and endocervical adenocarcinoma	179	272	259
Cholangiocarcinoma	35	34	33
Colon adenocarcinoma	255	275	219
Diffuse large B-cell lymphoma	%	32	32
Esophageal carcinoma (Esophageal adenocarcinoma + Esophageal squamous cell carcinoma)	179	75+78	73+77
Glioblastoma multiforme	143	140	138
Head and Neck squamous cell carcinoma	491	462	458
Kidney Chromophobe	65	65	64
Kidney renal clear cell carcinoma	423	342	342
Kidney renal papillary cell carcinoma	164	217	217
Acute Myeloid Leukemia	169	64	64
Brain Lower Grade Glioma	506	479	479
Liver hepatocellular carcinoma	187	336	333
Lung adenocarcinoma	487	418	393
Lung squamous cell carcinoma	174	449	433
Mesothelioma	%	76	76
Ovarian serous cystadenocarcinoma	341	248	248
Pancreatic adenocarcinoma	136	109	108
Pheochromocytoma and Paraganglioma	175	159	158

Prostate adenocarcinoma	409	423	421
Rectum adenocarcinoma	110	90	86
Sarcoma	237	185	183
Skin cutaneous melanoma	357	439	396
Stomach adenocarcinoma	335	327	279
Testicular Germ Cell Tumors	143	139	139
Thyroid carcinoma	393	240	240
Thymoma	116	72	71
Uterine Corpus Endometrial Carcinoma	229	373	335
Uterine Carcinosarcoma	53	53	52
Uveal melanoma	79	72	72
Total	**7828**	**7921**	**7626**

Table 7-3 Cell lines.

Cell lines were grown in indicated medium. Voltage, duration, and number of electroporation pulses are indicated for the CRISPR Cas9 nucleofection of individual gene pairs. Seeding density for proliferation assays are indicated for 96 well plates.

Cell line	Medium composition	Nucleofection conditions	Seeding density for proliferation assay [cells/96 well]
Hacat	DMEM+10% FCS	%	8000
HCC1973	RPMI+10% FCS	1550V, 10ms, 3 pulses	5000
HCT116	DMEM+10% FCS	1300V, 30ms, 1 pulse	3000
HEC1A	DMEM+10% FCS	1350V, 35ms, 2 pulses	7500
HEK293	DMEM+10% FCS	%	6000
JIMT1	DMEM+10% FCS	1550V, 10ms, 3 pulses	2000
KNS42	DMEM+10% FCS	1300V, 30ms, 1 pulse	8000
NCIH1975	RPMI+10% FCS	1300V, 10ms, 3 pulses	2000
NCIH2030	RPMI+10% FCS	1300V, 10ms, 3 pulses	2000
SF268	RPMI+10% FCS+1% glutamine	1300V, 30ms, 1 pulse	3000
TEN	1:1 DMEM:F-12 Nut Mix (Ham) +10% FCS	1350V, 35ms, 2 pulses	3750
U87MG	DMEM+10% FCS	1300V, 30ms, 1 pulse	3000

Table 7-4 Reagents, consumables and machines.

Reagents, consumables, and machines used for experiments, provider names, catalogue numbers and specifications are listed.

Reagent	Provider	Catalogue number	Specification
2-Mercaptoethanol	Sigma	M3148-2ml	
30X Reducing Agent (1.25M DTT)	Cell Signalling Technology	14265S	
3X Blue Loading Buffer	Cell Signalling Technology	56036S	
6well plate	Corning	CLS3516	6well cell culture plate, flat bottom with lid, tissue culture treated, non-pyrogenic, polysterene
96well plate	Thermo Fisher	167008	for proliferation assay; Nunc™ MicroWell™ 96-Well, Nunclon Delta-Treated, Flat-Bottom Microplate
96well plate	Greiner Bio-One	655161	for Incucyte® screen
Amersham™ ECL Western Blotting Analysis System	GE Healthcare	RPN2109	
Benzonase® Nuclease	Merck	E1014-25KU	
Crystal Violet	Bio Basic	CB0331	
Dimethyl Sulfoxide (DMSO)	Sigma	D2650-100ml	
Dried Skim Milk	Marvel		
Dulbecco's Modified Eagle's Medium (DMEM)	Sigma	D6429-500ml	high glucose with 4500mg/L glucose, L/glutamine, sodium pyruvate, and sodium bicarbonate, liquid, sterile-filtered, suitable for cell culture
Edit-R™ crRNA library combined with Edit-R™ tracrRNA	Horizon Discovery		custom library
F-12 Nut Mix (Ham)	Gibco	21765-029	+ L-Glutamine

Foetal Bovine Serum (FCS)	Labtech	FCS-SA 50115	
Formaldehyde	Thermo Fisher Scientific	F/1501/PB 17	Analytical Grade Reagent, 37-41% solution
FrameStar® 384	4ti-0385	4titude	
GenElute™ Mammalian Total RNA Miniprep kit	Sigma	RTN350-1KT	
Glutamine	Crick Media Preparation science technology platform		29.2g/l
gRNAs	Synthego		Table 7-6
High-Capacity cDNA Reverse Transcription Kit	Thermo Fisher	4368814	
Infinite® F200 Pro Plate Reader	Tecan		
INTERFERin®	Polyplus	409-01	
Incucyte® ZOOM	Sartorius		
Lipofectamine™2000 Transfection Reagent	Thermo Fisher	11668030	
Methanol	Thermo Fisher	M/3900/PB 17	
MiniAmp™ Thermal Cycler	Thermo Fisher	A37834	
Mini Trans-Blot® Filter paper	Bio-Rad	1703932	
Monarch® PCR and DNA Cleanup Kit (5µg)	New England BioLabs	T1030L	
Nancy-520	Merck	01494-500ul	
NanoDrop™ 1000 Spectrophotometer	Thermo Fisher		
Neon™ Transfection System	Thermo Fisher	MPK5000	
Neon™ Transfection System 10 µL Kit	Thermo Fisher	MPK1025	

Neubauer chamber	Brand GmbH+Co KG	717810	
Nuclease-free Water	Omega Bio-tek Inc.	NFWD062117SK3105	
O'GeneRuler™ 100bp DNA Ladder	Thermo Fisher	SM1143	
Opti-MEM™	Thermo Fisher	31985062	
OverNight Mix2Seq Kit	Eurofins Genomics		
PBS	Crick media preparation science technology platform		8g/l NaCl 0.25g/l KCl 1.437g/l Na_2HPO_4 0.25g/l KH_2PO_4 diluted in distilled water
PCR primers			see Table 7-5
Ponceau S Staining solution	Torcis	5225	
PowerUp™ SYBR™ Green Master Mix	Thermo Fisher	A25742	
Precision Plus Protein™ Dual Color Standards	Bio-Rad	1610374	
PureLink™ Genomic DNA Mini Kit	Thermo Fisher	K1820-02	
Q5® High Fidelity 2X Master Mix	New England BioLabs	M0492S	
QuantStudio™ Real-Time PCR Software (v1.2)	Thermo Fisher		
RNase Inhibitor	Thermo Fisher	N8080119	
Roswell Park Memorial Institute (RPMI) 1640 Medium	Sigma	R8758-500ml	with L-glutamine and sodium bicarbonate, liquid, sterile-filtered, suitable for cell culture
siGENOME siRNA pool library	Horizon Discovery		custom library
T25 cell culture flask	Corning	430168	

T75 cell culture flask	Thermo Fisher	156499	Nunc EasYFlask 75cm^2 Nunclon Delta Surface
TBS	Crick Media Preparation science technology platform		6.057g/l TRIS 8.766g/l NaCl 1l distilled water adjust to pH 7.5
TruPAGE™ Precast Gels 4-12%	Merck	PCG2011-10EA	
TruPAGE™ SDS Running Buffer, 20X	Merck	PCG3001-500ml	
TruPAGE™ Transfer Buffer, 20X	Merck	PCG3011-500ml	
Trypsin (10x)	Crick Media Preparation science technology platform		88.5g/l NaCl 2g/l KCl 11.5g/l Na_2HPO_4 2g/l KH_2PO_4 2g/l EDTA 5g/l Trypsin diluted in distilled water For working solution, dilute 1:10 in PBS
TrueCut™ Cas9 Protein v2 (5ug/µl)	Invitrogen	A36499	
Tween 20	Thermo Fisher	BP337-500	
UltraPure™ Agarose	Thermo Fisher	16500-500	
UVP BioDoc-It® UV Transilluminator	Analytik Jena		
ViiA™7 thermal cycler	Thermo Fisher	4453545	
X1 Amersham™ Protran™ 0.45µm Nitrocellulose membrane	GE Healthcare	10600002	

Table 7-5 PCR and qPCR primers.

Primers used for PCR and qPCR, their sequence and their provider are listed.

Primer	Sequence	Provider	Use
GATA2 forward	GGGTTGGCATAGTAGGGGTT	IDT	PCR
GATA2 reverse	CCGCCTTCCTTTCGTTTTGA	IDT	PCR
GATA3 forward	TTTGCTCACCTTTGCTTCCC	IDT	PCR
GATA3 reverse	CCTGACCGAGTTTCCGTAGT	IDT	PCR
MED13 forward	GCTGTTTCAAATAAAGTGGGCA	IDT	PCR
MED13 reverse	TGAACAGAGCACAGAACAAGT	IDT	PCR
MED13L forward	GGTTCATCTCCCCACCAGAA	IDT	PCR
MED13 reverse	ACATACATTCTTCACTGGGAGGA	IDT	PCR
MLLT1 forward	TCGTCGGCAGCGTCAGATGTGTATAGGA GACAGAACACCATCCAGTCGTGAGT	IDT	PCR
MLLT1 reverse	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGGTGGAATTGAGGATGAGGCG	IDT	PCR
MLLT3 forward	TCGTCGGCAGCGTCAGATGTGTATAAGA GACAGGCAGGGCTTGTGAAAGAGTC	IDT	PCR
MLLT3 reverse	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGTGTTTGATTGCCTGTGGTTCA	IDT	PCR
TBL1X forward	TCGTCGGCAGCGTCAGATGTGTATAAGA GACAGCCTGACCGGCCACTTTTAAT	IDT	PCR
TBL1X reverse	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGAGTAAGGAGAAAGACCGGGC	IDT	PCR
TBL1XR1 forward	TCGTCGGCAGCGTCAGATGTGTATAAGA GACAGCTGAATAAGCAAGCAAGACAGC	IDT	PCR
TBL1XR1 reverse	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGTGGGTAGGATGGGATTTGAAAA G	IDT	PCR
GAPDH forward	ACAGTTGCCATGTAGACC	Sigma-Aldrich	qPCR
GAPDH reverse	TTTTTGGTTGAGCACAGG	Sigma-Aldrich	qPCR
GATA2 forward	CTACTACAAGCTGCACAATG	Merck	qPCR
GATA2 reverse	CTTTCTTGCTCTTCTTGGAC	Merck	qPCR
GATA3 forward	AAAATGAACGGACAGAACC	Merck	qPCR

GATA3 reverse	GGGGTCTGTTAATATTGTGAAG	Merck	qPCR
MED13 forward	GGTGAATGGAAACAGTTCTATC	Merck	qPCR
MED13 reverse	AGCAACAAGAACTTCTACTG	Merck	qPCR
MED13L forward	CAAGTACTGGTAAGTCCTTATG	Merck	qPCR
MED13L reverse	CCATTCCTCAATCAACTTACG	Merck	qPCR
MLLT1 forward	CCAGCCATGGACAATCAGT	IDT	qPCR
MLLT1 reverse	GACAAACACCATCCAGTCGT	IDT	qPCR
MLLT3 forward	GAGCACAGTAACATACAGCACT	IDT	qPCR
MLLT3 reverse	ACCAGCATACCCAGATTCTTC	IDT	qPCR
TBL1X forward	CGGATGACATGACATTGAAG	Sigma-Aldrich	qPCR
TBL1X reverse	TCTTTATTGTGAGCCTGAAG	Sigma-Aldrich	qPCR
TBL1XR1 forward	CCAGAATATGGACTAAAGATGG	Sigma-Aldrich	qPCR
TBL1XR1 reverse	CTACTCCAGCACTTAGGATG	Sigma-Aldrich	qPCR

Table 7-6 gRNAs.

gRNAs used for CRISRP Cas9 knockout and their sequence are listed. All gRNAs were obtained from Synthego. Three gRNAs targeting a protein coding gene were mixed for each knockout, resulting in a final concentration of 2.38µM per gRNA. Two Negative Control Scrambled gRNAs were mixed for each control, resulting in a final concentration of 3.57µM per gRNA.

gRNA	Sequence
GATA2 #1	UGGCGCCGGAGCAGCCGCGC
GATA2 #2	GCCAUCCAGCGCGGCUGCUC
GATA2 #3	CUGCUCCGGCGCCACCUCCA
GATA3 #1	GAGCACAGCCGAGGCCAUGG
GATA3 #2	CUGGUCCGCCGUCACCUCCA
GATA3 #3	AGCCGAGGCCAUGGAGGUGA
MED13 #1	AGAAGCGACAGCUGCUAAAG
MED13 #2	AAGUAUUCUGCUUCAUCAGG
MED13 #3	AGGAAGUAUUCUGCUUCAUC
MED13L #1	UGAUUCCCGUGAGUUCAGCC
MED13L #2	UUUUUCCAGGCUGAACUCAC
MED13L #3	CUUUUUCCAGGCUGAACUCA
MLLT1 #1	ACCUCACCUGGACGGUGCAC
MLLT1 #2	UCCAGUGCACCGUCCAGGUG
MLLT1 #3	UCUCCUCCAGUGCACCGUCC
MLLT3 #1	GCAUUCUGAUGACAAUGAGG
MLLT3 #2	UCUGCAUUCUGAUGACAAUG
MLLT3 #3	UCUUUCAUUAUAGACCUCAA
TBL1X #1	CCUAGGCUGGUCCCUGCAGA
TBL1X #2	CGUCUGCAGGGACCAGCCUA
TBL1X #3	CCGUCUGCAGGGACCAGCCU
TBL1XR1 #1	AAGUUAGUAUUAAUGAGGUA
TBL1XR1 #2	GAUGAUAGAAAUCAAUGCAG
TBL1XR1 #3	GAUAUGGCUUUCUAUACCAA
Negative Control Scrambled (nTC) #1	GCACUACCAGAGCUAAC
Negative Control Scrambled (nTC) #2	GUACGUCGGUAUAACUC

Table 7-7 Antibodies.

Antibodies used for Western Blot, their targets, host animals in which they were produced, providers and dilution instructions are listed.

Target	Antibody name	Host	Provider	Dilution
Actin (42kDa)	A2103-200µl Anti Actin Sigma, Lot 115M4865V	rabbit	Sigma	1:2000 in 5% milk/TBST
MLLT1 (62kDa)	Anti-ENL antibody (ab49052)	rabbit	Abcam	7:5000 in 5% milk/TBST
H3 (15kDa)	Histone H3 (D1HH2) XP mAb #4499	rabbit	Cell Signalling Technology	1:1000 in 5% milk/TBST
H3K79me2 (15kDa)	Di-Methyl-Histone H3 (Lys79) (D15E8) XP mAb #5427	rabbit	Cell Signalling Technology	1:1000 in 5% milk/TBST
H3K79me3 (15kDa)	Recombinant Anti-Histone H3 (tri methyl K79) antibody, ab208189	rabbit	Abcam	1:1000 in 5% milk/TBST
Anti-rabbit	Anti-Rabbit IgG, HRP-linked Antibody (#7074)	goat	Cell Signalling Technology	1:5000 in 5% milk/TBST

Table 7-8 ENTREZ IDs of 279 chromatin modifiers, 51 DNA modifiers and 685 histone modifiers.

Chromatin modifiers

60, 86, 142, 546, 605, 648, 676, 892, 1024, 1060, 1106, 1107, 1108, 1616, 1663, 1911, 1912, 2074, 2186, 2308, 2353, 2354, 2355, 2491, 2908, 3146, 3148, 3149, 3150, 3151, 3159, 3169, 3170, 3171, 3297, 3725, 3726, 3727, 3930, 4001, 4005, 4087, 4149, 4171, 4609, 4641, 4673, 4674, 4675, 4676, 4678, 4691, 4798, 4869, 5079, 5469, 5496, 5499, 5500, 5501, 5514, 5885, 5928, 5931, 6015, 6045, 6046, 6294, 6322, 6418, 6594, 6595, 6596, 6597, 6598, 6599, 6601, 6602, 6603, 6604, 6605, 6749, 6760, 6830, 6837, 6941, 6944, 7141, 7142, 7150, 7153, 7155, 7270, 7290, 7528, 7703, 7913, 8061, 8099, 8193, 8208, 8243, 8289, 8295, 8438, 8451, 8467, 8535, 8607, 8861, 9031, 9112, 9126, 9219, 9274, 9275, 9282, 9324, 9412, 9439, 9440, 9441, 9442, 9443, 9477, 9557, 9631, 9640, 9688, 9862, 9878, 9918, 9968, 9969, 9972, 10025, 10036, 10051, 10138, 10155, 10274, 10320, 10336, 10361, 10389, 10445, 10514, 10524, 10541, 10592, 10661, 10664, 10734, 10735, 10743, 10847, 10856, 10902, 10951, 11177, 11198, 11335, 11339, 22806, 22893, 22955, 22985, 23347, 23389, 23394, 23397, 23414, 23421, 23429, 23466, 23468, 23476, 23492, 23523, 23613, 23636, 25788, 25842, 26013, 26038, 26039, 26040, 27000, 27086, 27443, 29117, 29844, 29855, 30836, 50485, 50511, 50809, 51003, 51377, 51412, 51460, 51586, 51773, 54069, 54107, 54108, 54556, 54617, 54797, 54815, 54891, 55090, 55166, 55193, 55205, 55274, 55320, 55355, 55588, 55636, 55723, 55839, 56916, 57332, 57459, 57492, 57504, 57592, 57634, 57666, 57680, 64105, 64151, 64319, 64946, 78994, 79019, 79172, 79366, 79682, 79913, 80012, 80152, 80205, 80306, 80335, 81611, 81857, 83444, 83746, 83983, 84108, 84181, 84246, 84295, 84333, 84619, 84733, 84759, 85509, 90390, 91687, 93973, 93986, 112950, 116113, 116447, 116931, 124944, 125476, 148479, 196528, 201161, 219541, 266812, 283899, 378708, 400569, 401541

DNA modifiers

328, 339, 1060, 1647, 1660, 1786, 1787, 1788, 1789, 3070, 4152, 4204, 4255, 4300, 4616, 5111, 6688, 6996, 8099, 8846, 8930, 8932, 10009, 10664, 10912, 10919, 10930, 11176, 23512, 26097, 27350, 29128, 29947, 30827, 53615, 54737, 54790, 57379, 57659, 79727, 79813, 80312, 84944, 115426, 121642, 140690, 200315, 200424, 221120, 253461, 346171

Histone modifiers

60, 86, 142, 322, 326, 406, 408, 429, 473, 545, 546, 579, 580, 605, 639, 648, 672, 676, 686, 699, 983, 1017, 1018, 1020, 1024, 1025, 1104, 1105, 1107, 1108, 1111, 1147, 1207, 1326, 1385, 1386, 1387, 1482, 1487, 1488, 1613, 1620, 1642, 1643, 1655, 1660, 1810, 1820, 1843, 1876, 1911, 1912, 2033, 2070, 2078, 2091, 2099, 2122, 2138, 2139, 2140, 2145, 2146, 2186, 2332, 2534, 2624, 2625, 2648, 2672, 2931, 2932, 2956, 2969, 3054, 3065, 3066, 3141, 3150, 3169, 3187, 3275, 3276, 3297, 3303, 3619, 3621, 3622, 3717, 3720, 3832, 3984, 3985, 4086, 4087, 4088, 4090, 4093, 4094, 4149, 4152, 4204, 4221, 4297, 4298, 4300, 4302, 4435, 4436, 4485, 4591, 4602, 4609, 4613, 4641, 4674, 4691, 4734, 4796, 4862, 4998, 4999, 5000, 5058, 5062, 5079, 5087, 5128, 5245, 5252, 5253, 5315, 5496, 5499, 5500, 5501, 5531, 5537, 5562, 5563, 5564, 5565, 5571, 5579, 5580, 5585, 5591, 5896, 5897, 5914, 5926, 5927, 5928, 5929, 5931, 5933, 5934, 5977, 5978, 5987, 6015, 6045, 6046, 6197, 6304, 6314, 6322, 6419, 6421, 6497, 6500, 6591, 6606, 6621, 6622, 6662, 6667, 6672, 6688, 6722, 6789, 6790, 6793, 6795, 6830, 6839, 6871, 6872, 6873, 6874, 6877, 6878, 6879, 6880, 6881, 6883, 6885, 6895, 6907, 6908, 6941, 6944, 7027, 7088, 7091, 7158, 7284, 7291, 7317, 7319, 7320, 7323, 7324, 7328, 7334, 7403, 7404, 7443, 7468, 7490, 7528, 7552, 7702, 7703, 7704, 7750, 7764, 7799, 7862, 7874, 7994, 8019, 8028, 8085, 8087, 8089, 8091, 8099, 8110, 8125, 8202, 8237, 8241, 8242, 8284, 8295, 8314, 8328, 8405, 8450, 8451, 8464, 8467, 8473, 8479, 8493, 8505, 8518, 8520, 8535, 8607, 8626, 8648, 8726, 8805, 8819, 8841, 8850, 8861, 8932, 8986, 9025, 9031, 9070, 9085, 9112, 9202, 9203, 9212, 9219, 9252, 9329, 9425, 9513, 9541, 9575, 9577, 9611, 9612, 9640, 9643, 9646, 9656,

9658, 9666, 9678, 9682, 9701, 9733, 9734, 9739, 9757, 9759, 9767, 9782, 9810, 9866, 9869, 9874, 9913, 9958, 9960, 9978, 9989, 10013, 10014, 10038, 10039, 10043, 10075, 10138, 10155, 10284, 10320, 10322, 10336, 10362, 10363, 10370, 10389, 10419, 10432, 10445, 10474, 10498, 10514, 10521, 10524, 10600, 10620, 10626, 10629, 10664, 10765, 10771, 10783, 10856, 10891, 10902, 10919, 10927, 10933, 10943, 10951, 11083, 11091, 11105, 11107, 11108, 11143, 11168, 11176, 11240, 11262, 11329, 11335, 22806, 22823, 22870, 22933, 22955, 22976, 22992, 23013, 23028, 23030, 23040, 23054, 23067, 23081, 23126, 23133, 23135, 23186, 23199, 23210, 23243, 23269, 23272, 23304, 23309, 23314, 23326, 23338, 23347, 23378, 23394, 23408, 23409, 23410, 23411, 23414, 23424, 23429, 23450, 23466, 23468, 23476, 23492, 23512, 23515, 23522, 23569, 23587, 23598, 23613, 23774, 25836, 25855, 25862, 25936, 25942, 25988, 26009, 26013, 26038, 26040, 26097, 26108, 26122, 26147, 26155, 26168, 26610, 27000, 27005, 27043, 27044, 27086, 27097, 27154, 29028, 29072, 29086, 29117, 29128, 29915, 29943, 29947, 29994, 30836, 50943, 51105, 51111, 51132, 51147, 51176, 51230, 51317, 51322, 51422, 51460, 51533, 51535, 51547, 51548, 51562, 51564, 51592, 51616, 51720, 51742, 51773, 51780, 53615, 53632, 54014, 54093, 54107, 54454, 54496, 54531, 54556, 54623, 54625, 54704, 54737, 54790, 54799, 54815, 54859, 54880, 54881, 54904, 54934, 54971, 55023, 55140, 55148, 55167, 55170, 55205, 55209, 55226, 55250, 55252, 55257, 55291, 55578, 55662, 55671, 55683, 55689, 55693, 55729, 55758, 55777, 55791, 55806, 55818, 55869, 55870, 55872, 55904, 55929, 56254, 56341, 56655, 56852, 56916, 56943, 56946, 56950, 56970, 56978, 56979, 56981, 57117, 57215, 57223, 57325, 57332, 57459, 57504, 57592, 57602, 57634, 57649, 57661, 57666, 57680, 57708, 57713, 57718, 57798, 58487, 58508, 59335, 59336, 63035, 63925, 63976, 63978, 64061, 64319, 64324, 64426, 64754, 64769, 64854, 65980, 79084, 79142, 79184, 79447, 79577, 79595, 79685, 79697, 79718, 79723, 79813, 79831, 79885, 79903, 79918, 79960, 80012, 80063, 80204, 80312, 80314, 80335, 80816, 80853, 80854, 81550, 83746, 83852, 83860, 83903, 83933, 84101, 84108, 84142, 84148, 84159, 84193, 84215, 84289, 84295, 84312, 84333, 84444, 84456, 84525, 84619,

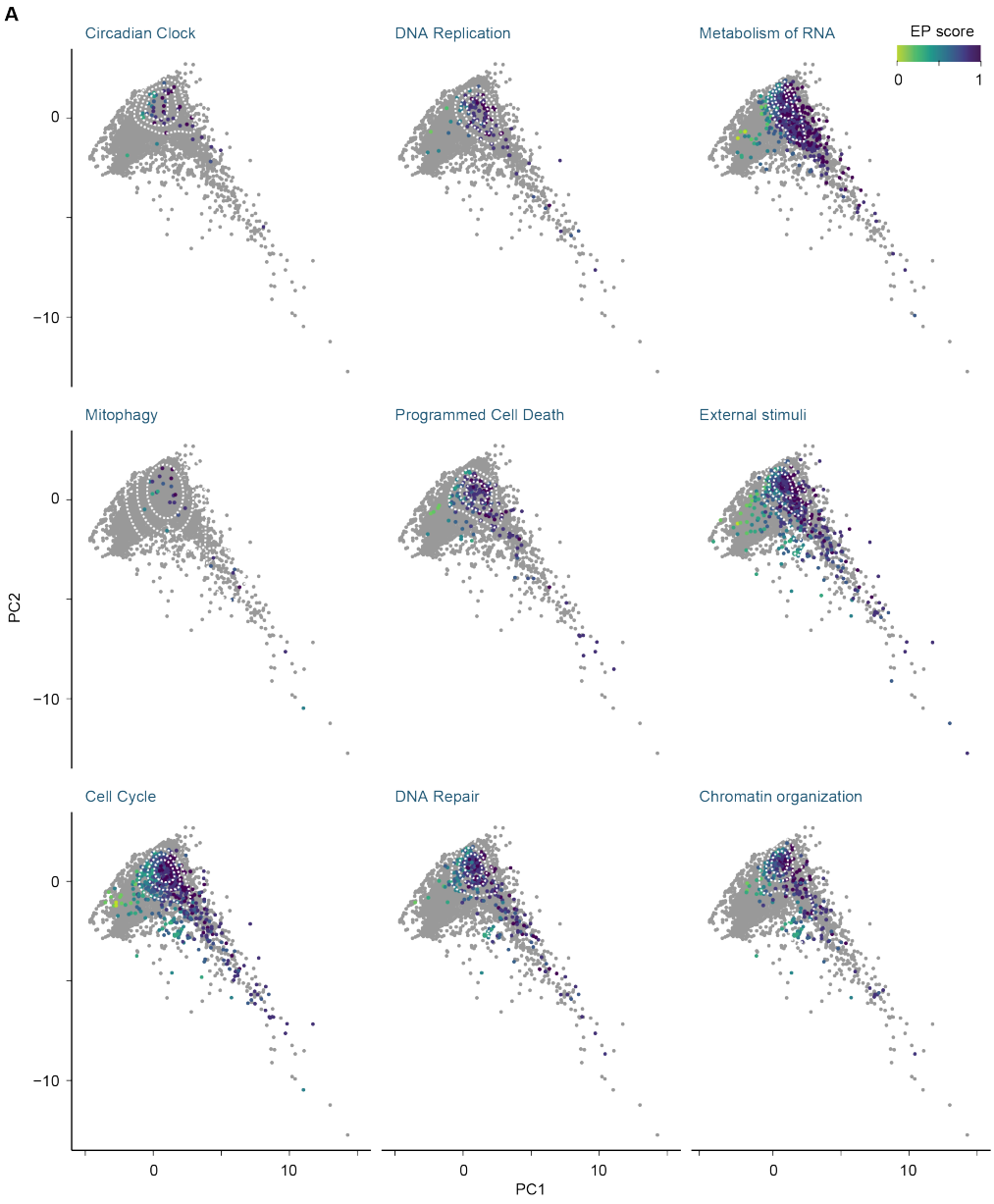
84656, 84661, 84678, 84717, 84733, 84759, 84787, 84844, 84864, 84875, 84930, 84939, 85509, 90780, 91526, 91754, 93166, 93624, 93649, 93973, 93986, 112869, 112970, 114785, 114803, 114825, 115426, 116113, 117143, 121536, 122953, 123169, 124359, 124944, 127002, 127540, 129685, 138474, 140690, 148479, 150572, 151636, 151987, 152098, 157313, 163732, 165918, 171023, 200424, 219333, 221656, 222229, 222255, 253175, 257218, 260434, 283248, 283373, 284058, 286204, 339287, 346171, 359787, 377630, 387893, 390245, 100137047, 100316904

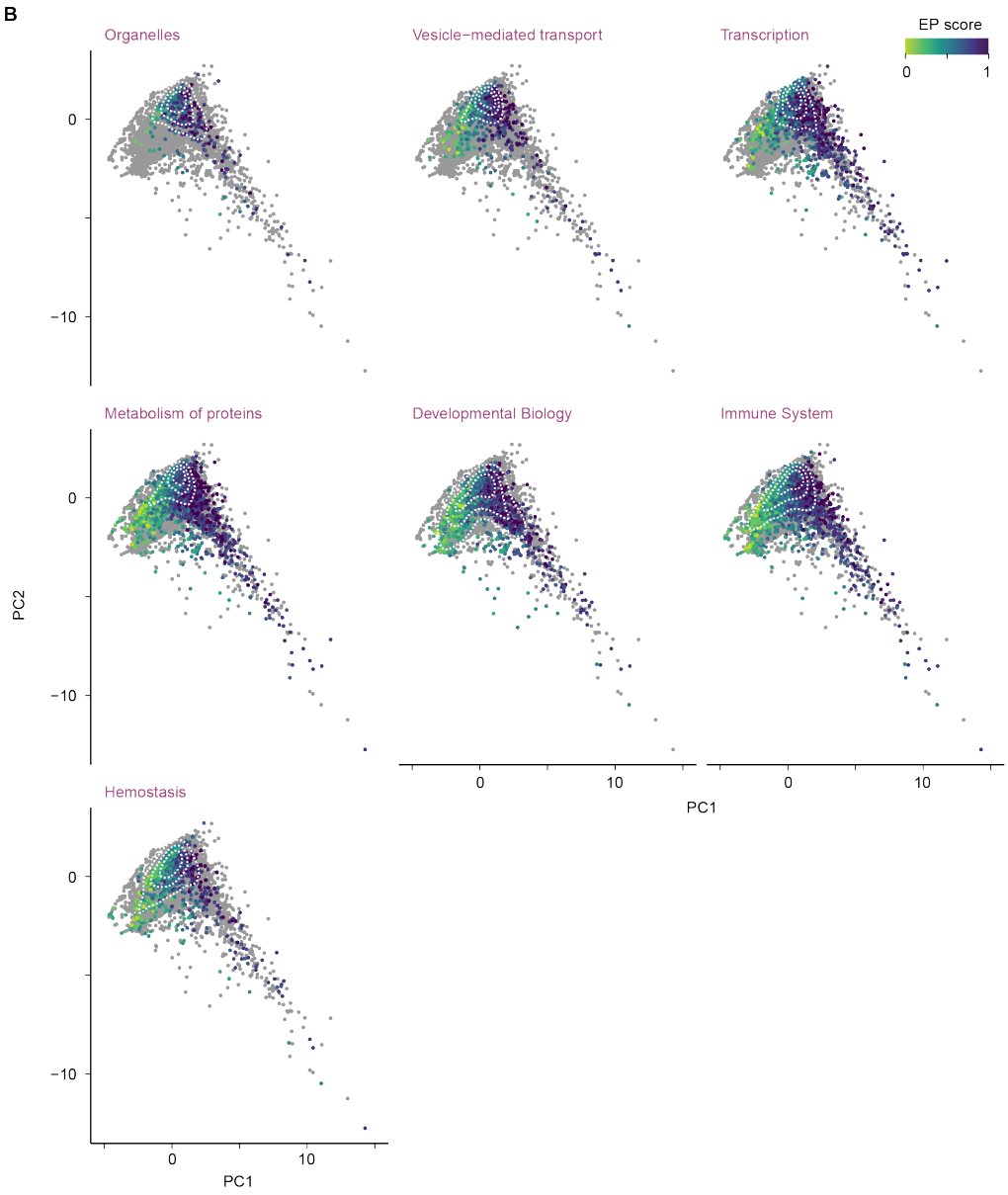
Table 7-9 List of predicted functional compensator pairs from the improved prediction approach.

Fifty-four gene pairs were predicted as potential functional compensators by the improved prediction approach. Gene symbols for both partners are listed, ordered according to the sum of samples in TCGA with a damaging alteration in gene A or gene B. The decision which gene was gene A was based on alphabetical order, known synthetic lethal partners are indicated with a *. Involvement in epigenetics is indicated by “EM”. Adjusted p-values for a decreased expression (expr.) of B when A is essential (ess.) in cancer cell lines, and vice versa, are indicated. Gene pairs were included in the list if either p value was below 0.1. ARID1A/ARID1B and CREBBP/EP300 were included because of significantly higher essentiality of one gene when the other had a damaging alteration.

gene A	gene B	EM status	damaged samples A (n)	damaged samples B (n)	damaged samples total (n)	p (A ess., B expr.)	p (B ess., A expr.)
ARID1A*	ARID1B*	EM	518	157	675	0.89	0.28
CREBBP*	EP300*	EM	273	233	506	0.22	1
SMARCA2*	SMARCA4*	EM	120	200	320	5.70E-05	2.40E-05
MED13	MED13L	EM	110	92	202	1	0.073
STAG1*	STAG2*	EM	68	132	200	2.30E-07	2.40E-07
GATA2	GATA3	EM	38	131	169	0.019	1
ITSN1	ITSN2		102	44	146	NA	0.053
PPP2R1A	PPP2R1B		105	32	137	8.30E-16	NA
PDS5B	PDS5A		78	51	129	0.31	0.012
PICALM	SNAP91		42	76	118	0.045	NA
ACO1	IREB2		55	58	113	NA	7.00E-04
ELMO1	ELMO2		77	22	99	0.97	9.00E-06
RNF40	RNF20	EM	45	46	91	0.011	1
SEC24B	SEC24A		47	41	88	0.042	NA
TBL1X	TBL1XR1	EM	26	56	82	NA	1.20E-05
PRKAA1	PRKAA2	EM	26	54	80	0.0012	0.17
CAB39	CAB39L		10	70	80	0.0016	NA
ARNT	ARNT2		37	41	78	0.04	NA
SPTLC2	SPTLC3		27	49	76	0.00066	NA
CUL4B	CUL4A	EM	51	21	72	0.0027	0.057
DDX5	DDX17	EM	26	46	72	1.20E-05	0.0032
BTRC	FBXW11		38	29	67	NA	3.90E-12
RNF19A	RNF19B		50	17	67	0.012	NA
VPS4B	VPS4A		35	30	65	1.40E-09	1.40E-08
CDK8	CDK19	EM	36	27	63	0.057	NA
FNIP2	FNIP1		29	29	58	0.26	0.042
CSTF2	CSTF2T		27	24	51	0.00045	0.18
PSEN1	PSEN2		19	31	50	0.036	NA
AP1M1	AP1M2		28	18	46	0.006	1
ATP6V0D1	ATP6V0D2		17	29	46	0.057	NA
TUBG1	TUBG2		24	19	43	5.30E-08	0.22
DYNC1LI2	DYNC1LI1		17	25	42	0.067	0.35
ETS1	ETS2		20	21	41	NA	0.011
NDE1	NDEL1		27	14	41	6.10E-06	NA
RING1	RNF2	EM	20	20	40	0.0016	NA
SEC61A1	SEC61A2		21	19	40	0.00056	0.38
EAF2	EAF1		29	10	39	0.46	5.60E-16
COPG1	COPG2		38	0	38	2.90E-16	NA
RRAGD	RRAGC		21	12	33	NA	0.00012
DDX39B	DDX39A		18	13	31	9.00E-06	0.26
PRKAB1	PRKAB2	EM	18	11	29	0.029	NA
TIMM17B	TIMM17A		14	12	26	NA	3.10E-11
ASF1A	ASF1B	EM	12	14	26	0.032	9.40E-06
VAPB	VAPA		12	13	25	NA	0.04
CDK4	CDK6		15	9	24	1.30E-17	0.00011
BCL2	BCL2L1		15	7	22	1.00E-09	3.10E-05
NAPA	NAPB		7	15	22	6.40E-06	NA
NABP1	NABP2		12	9	21	NA	0.011
CRK	CRKL		7	13	20	0.00066	0.69
DERL2	DERL3		10	9	19	0.087	NA
CHMP4C	CHMP4B		8	11	19	0.55	2.70E-46
CDK11A	CDK11B		8	0	8	0.0091	NA
DYNLL1	DYNLL2		4	1	5	5.30E-08	0.14
RPL26	RPL26L1		2	2	4	0.067	0.8

7.2 Supplementary Figures




B
Organelles
Vesicle−mediated transport
Transcription
EP score
0
1
Metabolism of proteins
Developmental Biology
Immune System
Hemostasis
PC2
PC1
0
10
−10

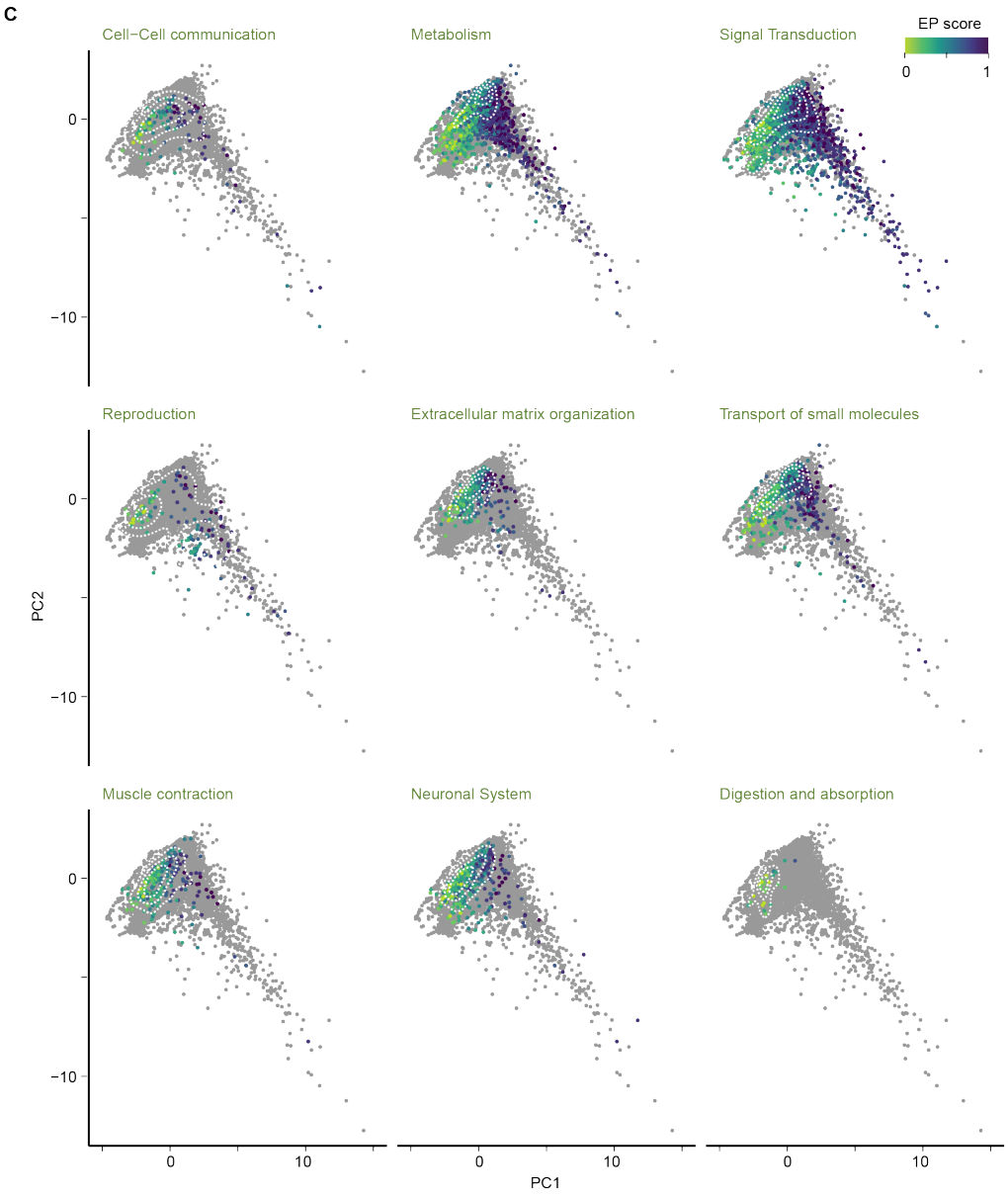
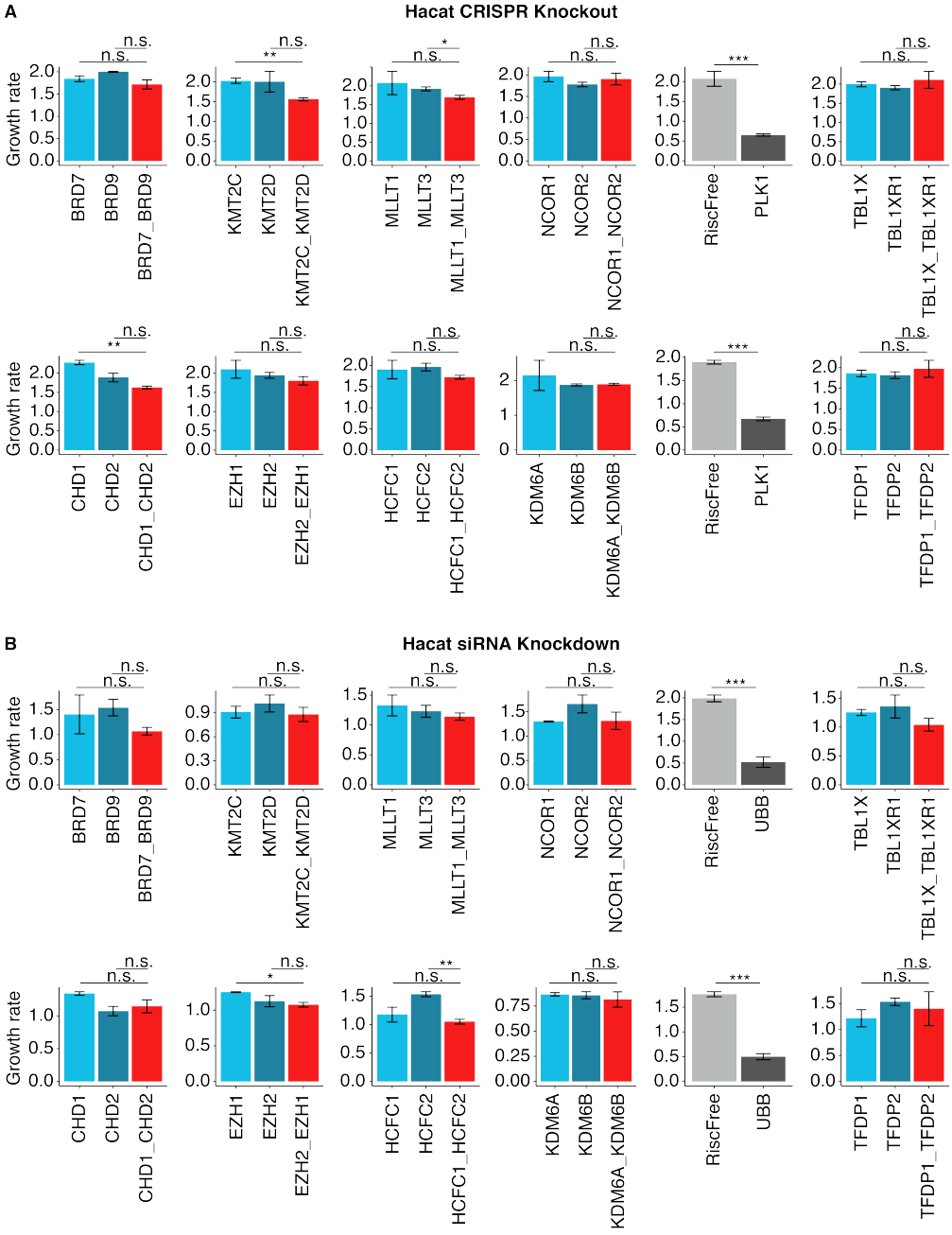


Figure 7-1 Position of pathway-specific genes in the Principal Component Analysis.

PCA including nine evolutionary properties and 10,291 Reactome genes. The score plot shows contributions of the first and second principal components, each dot represents one gene. Forty-three outlier genes with PPIN degree >500 were removed. Genes belonging to the respective **A)** group 1, **B)** group 2 and **C)** group 3 pathway are highlighted in colour representing their EP score. Density lines are indicated in white.


A
Hacat CRISPR Knockout
Growth rate
BRD7
BRD9
BRD7_BRD9
KMT2C
KMT2D
KMT2C_KMT2D
MLLT1
MLLT3
MLLT1_MLLT3
NCOR1
NCOR2
NCOR1_NCOR2
RiscFree
PLK1
TBL1X
TBL1XR1
TBL1X_TBL1XR1
CHD1
CHD2
CHD1_CHD2
EZH1
EZH2
EZH2_EZH1
HCFC1
HCFC2
HCFC1_HCFC2
KDM6A
KDM6B
KDM6A_KDM6B
TFDP1
TFDP2
TFDP1_TFDP2
n.s.
B
Hacat siRNA Knockdown
UBB

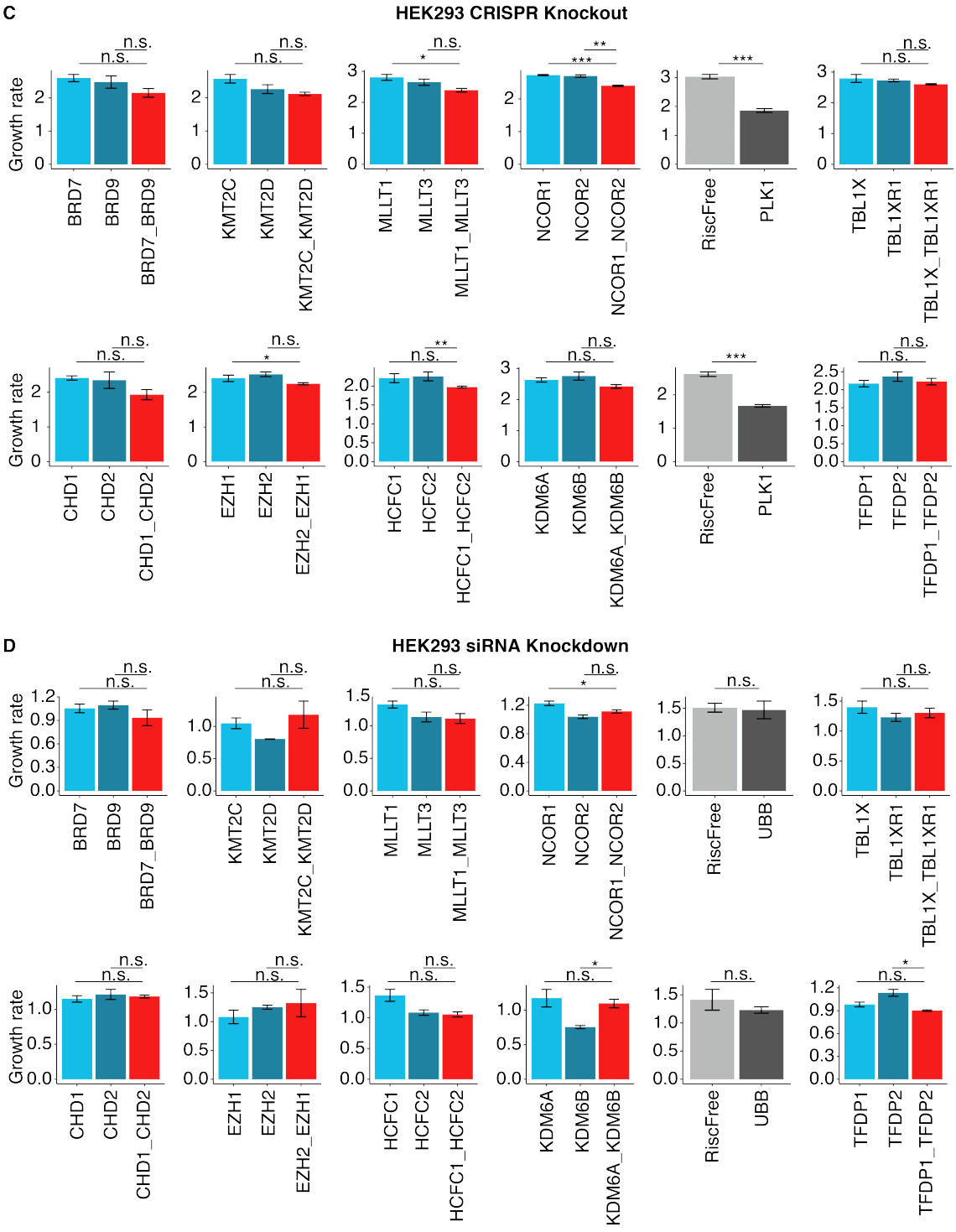


Figure 7-2 Growth rates of CRISPR Cas9 knockout and siRNA knockdown screens. Growth rates in the exponential growth phase of single and double gene **A,C)** CRISPR Cas9 knockout and **B,D)** siRNA knockdown in A,B) Hacat and C,D) HEK293 cells. A two-sided t test was used to compare growth rates between single and double gene knockdown/ knockout. n.s. – non significant, * p<0.05, ** p<0.01, *** p<0.001. Despite the seemingly similar growth rates of HEK293

siRNA positive and negative control, the knockdown of *UBB* resulted in cell death as shown in Figure 5-5, and the observed high growth rate was calculated in the first few hours of the experiment before cell death occurred.

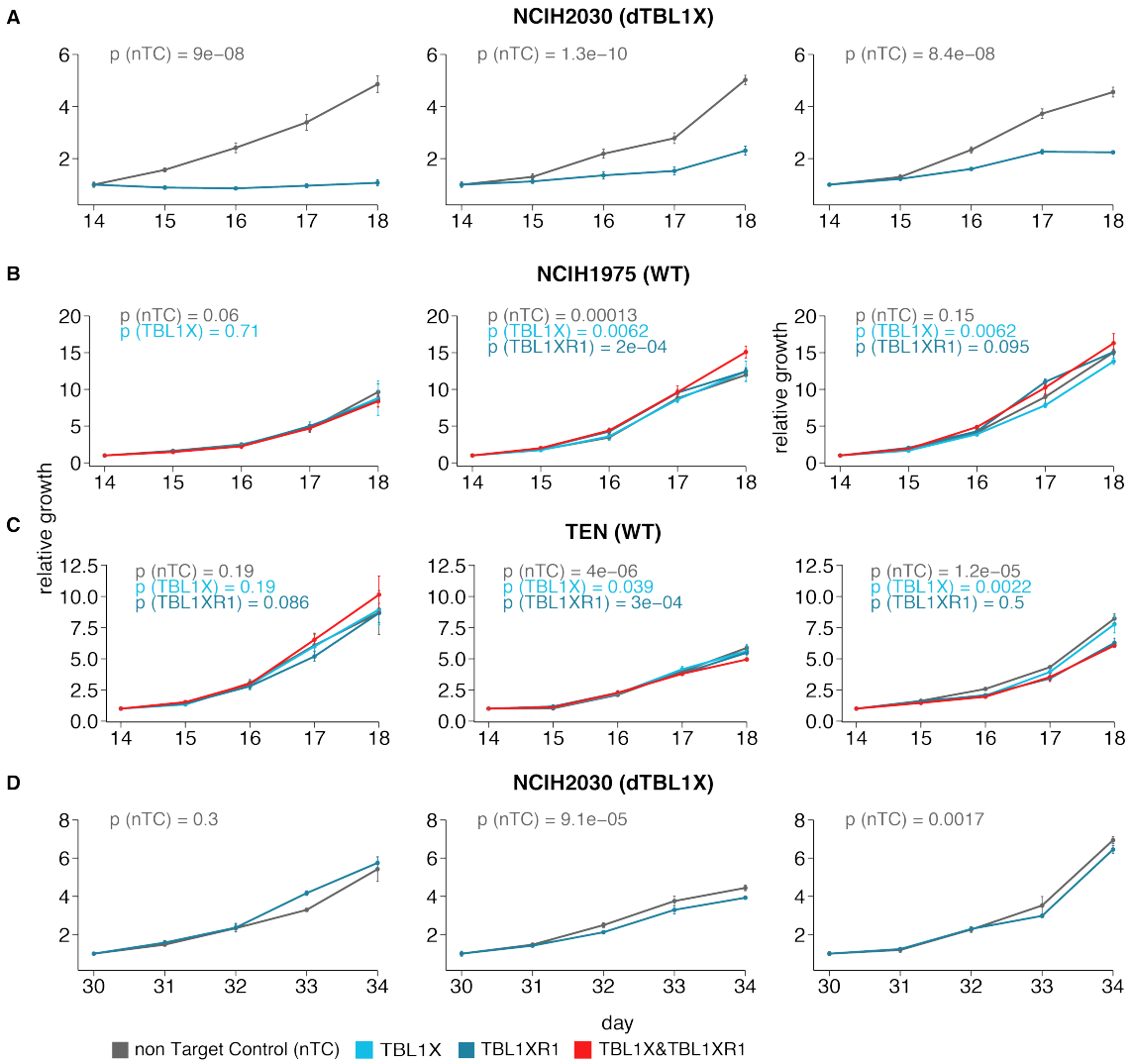


Figure 7-3 Proliferation of cell lines with *TBL1X*, *TBL1XR1* or double gene knockout.

Recombinant Cas9 protein and sgRNAs were introduced into **A,D)** NCIH2030, **B)** NCIH1975 and **C)** TEN cells using nucleofection. After A,B,C) 14 or D) 30 days, proliferation assays were started and run for four days. Three independent repeats are shown for each cell line. P values were calculated using a two-sided t test and indicate significance at the last day of the assay between A,D) nTC and *TBL1XR1* knockout cells or between **B,C)** nTC, *TBL1X* or *TBL1XR1* knockout and the double knockout.

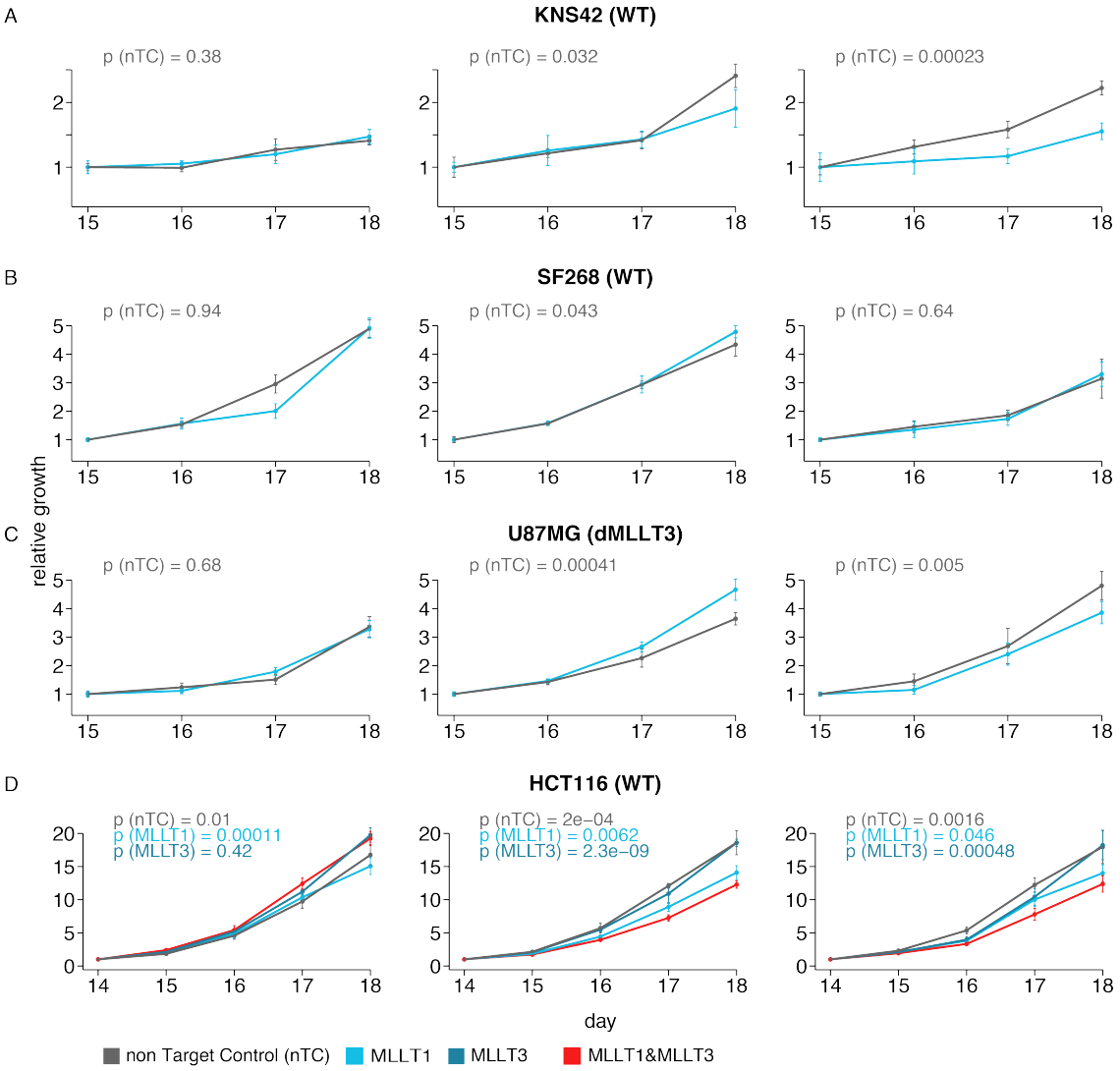


Figure 7-4 Proliferation of cell lines with *MLLT1*, *MLLT3* or double gene knockout. Recombinant Cas9 protein and sgRNAs were introduced into **A)** KNS42, **B)** SF268, **C)** U87MG and **D)** HCT116 cells using nucleofection. After 14 days, proliferation assays were started and run for A,B,C) four or D) seven days. Three independent repeats are shown for each cell line. P values were calculated using a two-sided t test and indicate significance at the last day of the assay between A,B,C) nTC and *MLLT1* knockout cells or between D) nTC, *MLLT1* or *MLLT3* knockout and the double knockout.

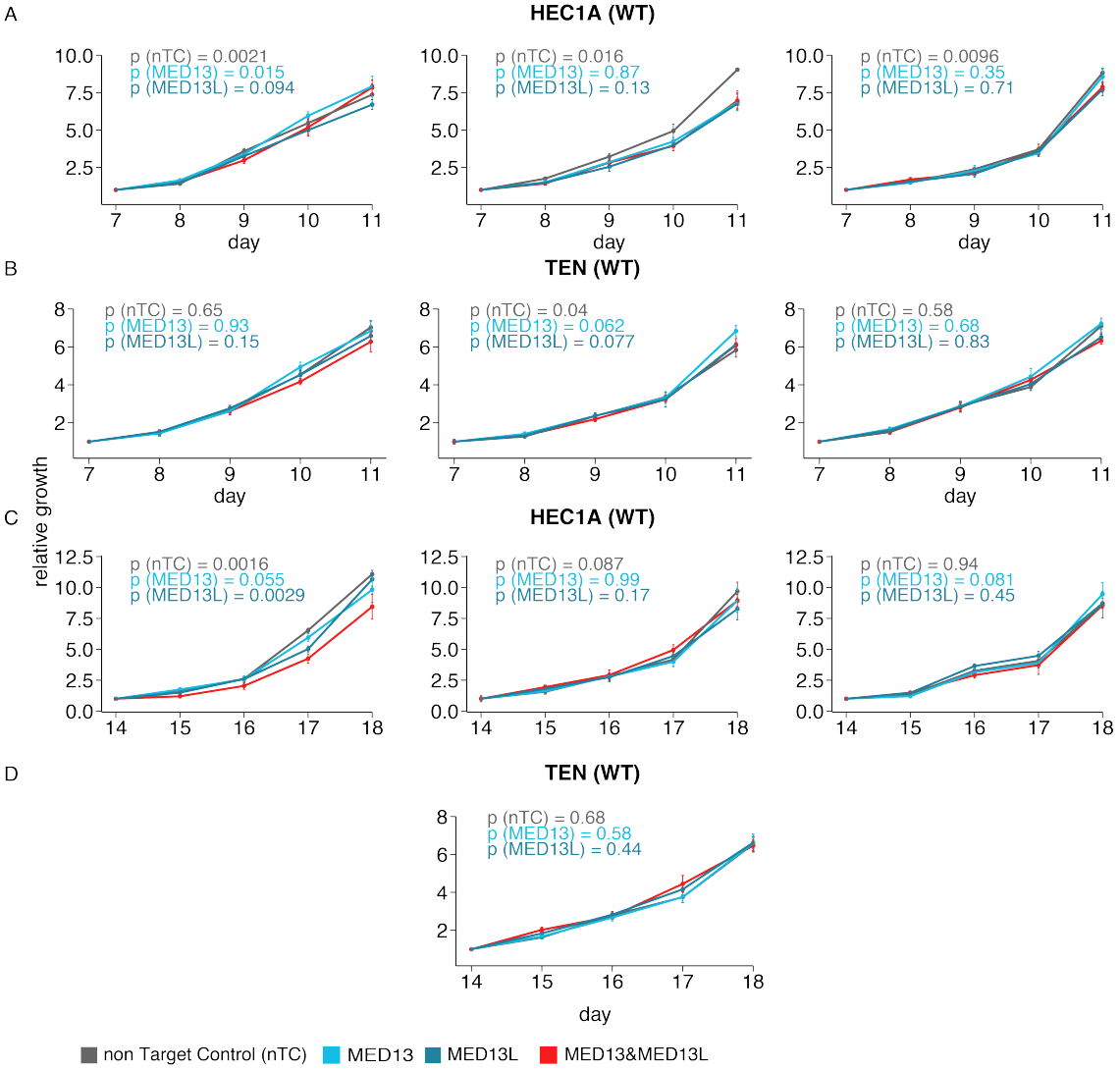


Figure 7-5 Proliferation of cell lines with *MED13*, *MED13L* or double gene knockout. Recombinant Cas9 protein and sgRNAs were introduced into **A,C)** HEC1A and **B,D)** TEN cells using nucleofection. Proliferation was assessed A,B) seven or C,D) 14 days after nucleofection for four days. Three independent repeats are shown for each cell line, except for figure (D) where the experiment was only repeated once. P values were calculated using a two-sided t test and indicate significance at the last day of the assay between nTC, *MED13* or *MED13L* knockout and the double knockout.

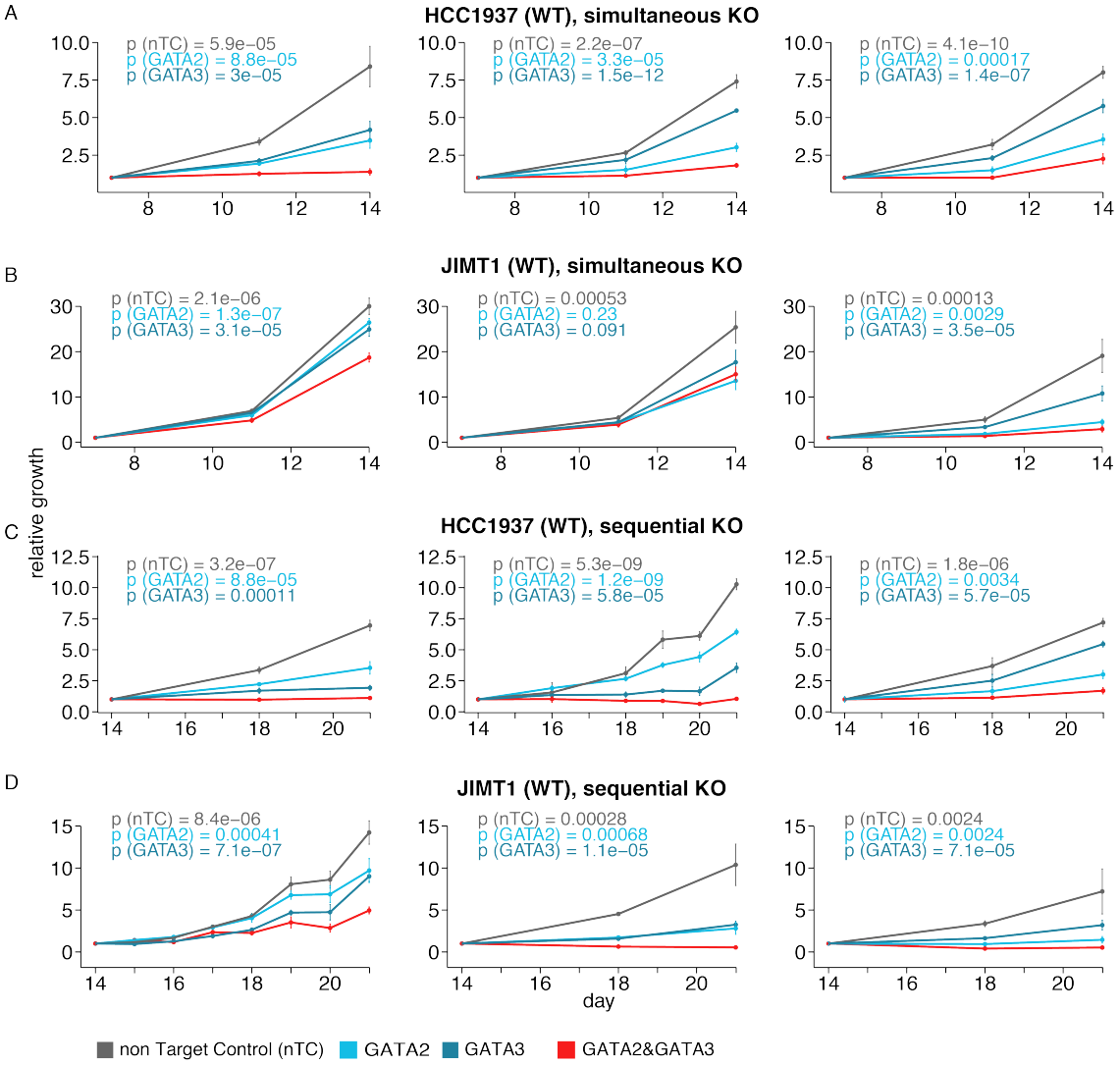


Figure 7-6 Proliferation of cell lines with *GATA2*, *GATA3* or double gene knockout. Recombinant Cas9 protein and sgRNAs were introduced into **A,C)** HCC1937 and **B,D)** JIMT1 cells using nucleofection. Double gene knockout was achieved as A,B) a simultaneous knockout or as C,D) a sequential knockout of *GATA2* and after seven days *GATA3*. Seven days after the last knockout, proliferation assays were started and run for seven days. Three independent repeats are shown for each cell line. P values were calculated using a two-sided t test and indicate significance at the last day of the assay between nTC, *GATA2* or *GATA3* knockout and the double knockout.

References

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet *Chapter 7*, Unit7 20. 10.1002/0471142905.hg0720s76.

Agarwal, R., Kumar, B., Jayadev, M., Raghav, D., and Singh, A. (2016). CoReCG: a comprehensive database of genes associated with colon-rectal cancer. Database (Oxford) *2016*. 10.1093/database/baw059.

Ainscough, B.J., Griffith, M., Coffman, A.C., Wagner, A.H., Kunisaki, J., Choudhary, M.N., McMichael, J.F., Fulton, R.S., Wilson, R.K., Griffith, O.L., and Mardis, E.R. (2016). DoCM: a database of curated mutations in cancer. Nat Methods *13*, 806-807. 10.1038/nmeth.4000.

Akcakaya, P., Bobbin, M.L., Guo, J.A., Malagon-Lopez, J., Clement, K., Garcia, S.P., Fellows, M.D., Porritt, M.J., Firth, M.A., Carreras, A., et al. (2018). In vivo CRISPR editing with no detectable genome-wide off-target mutations. Nature *561*, 416-419. 10.1038/s41586-018-0500-9.

Albert, R., Jeong, H., and Barabasi, A.L. (2000). Error and attack tolerance of complex networks. Nature *406*, 378-382. 10.1038/35019019.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2015). Molecular Biology of the Cell, 6th Edition (Garland Science).

Alkan, F., Wenzel, A., Anthon, C., Havgaard, J.H., and Gorodkin, J. (2018). CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. Genome Biol *19*, 177. 10.1186/s13059-018-1534-x.

Ameratunga, M., Xu, W., and Lopez, J. (2020). Personalized Cancer Immunotherapy: Today's Challenge and Tomorrow's Promise. Journal of Immunotherapy and Precision Oncology *1*, 56-67. 10.4103/jipo.Jipo_13_18.

An, O., Dall'Olio, G.M., Mourikis, T.P., and Ciccarelli, F.D. (2016). NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. Nucleic Acids Res *44*, D992-999. 10.1093/nar/gkv1123.

An, O., Pendino, V., D'Antonio, M., Ratti, E., Gentilini, M., and Ciccarelli, F.D. (2014). NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database (Oxford) *2014*, bau015. 10.1093/database/bau015.

Anastasiadi, D., Esteve-Codina, A., and Piferrer, F. (2018). Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics Chromatin *11*, 37. 10.1186/s13072-018-0205-1.

Anglesio, M.S., Papadopoulos, N., Ayhan, A., Nazeran, T.M., Noë, M., Horlings, H.M., Lum, A., Jones, S., Senz, J., Seckin, T., et al. (2017). Cancer-Associated Mutations in Endometriosis without Cancer. New England Journal of Medicine *376*, 1835-1848. 10.1056/NEJMoa1614814.

Aran, D., Toperoff, G., Rosenberg, M., and Hellman, A. (2011). Replication timing-related and gene body-specific methylation of active human genes. Hum Mol Genet *20*, 670-680. 10.1093/hmg/ddq513.

Aster, J.C., Pear, W.S., and Blacklow, S.C. (2017). The Varied Roles of Notch in Cancer. Annu Rev Pathol *12*, 245-275. 10.1146/annurev-pathol-052016-100127.

Bailey, A.M., Mao, Y., Zeng, J., Holla, V., Johnson, A., Brusco, L., Chen, K., Mendelsohn, J., Routbort, M.J., Mills, G.B., and Meric-Bernstam, F. (2014). Implementation of biomarker-driven cancer therapy: existing tools and remaining gaps. Discov Med *17*, 101-114.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive

Characterization of Cancer Driver Genes and Mutations. Cell *174*, 1034-1035. 10.1016/j.cell.2018.07.034.
Bajrami, I., Marlow, R., van de Ven, M., Brough, R., Pemberton, H.N., Frankum, J., Song, F., Rafiq, R., Konde, A., Krastev, D.B., et al. (2018). E-Cadherin/ROS1 Inhibitor Synthetic Lethality in Breast Cancer. Cancer Discov *8*, 498-515. 10.1158/2159-8290.CD-17-0603.
Barth, T.K., and Imhof, A. (2010). Fast signals and slow marks: the dynamics of histone modifications. Trends Biochem Sci *35*, 618-626. 10.1016/j.tibs.2010.05.006.
Bartha, I., di Iulio, J., Venter, J.C., and Telenti, A. (2018). Human gene essentiality. Nat Rev Genet *19*, 51-62. 10.1038/nrg.2017.75.
Behan, F.M., Iorio, F., Picco, G., Goncalves, E., Beaver, C.M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., et al. (2019). Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature *568*, 511-516. 10.1038/s41586-019-1103-9.
Benedetti, L., Cereda, M., Monteverde, L., Desai, N., and Ciccarelli, F.D. (2017). Synthetic lethal interaction between the tumour suppressor STAG2 and its paralog STAG1. Oncotarget *8*, 37619-37632. 10.18632/oncotarget.16838.
Benstead-Hume, G., Chen, X., Hopkins, S.R., Lane, K.A., Downs, J.A., and Pearl, F.M.G. (2019). Predicting synthetic lethal interactions using conserved patterns in protein interaction networks. PLoS Comput Biol *15*, e1006888. 10.1371/journal.pcbi.1006888.
Berns, K., Hijmans, E.M., Mullenders, J., Brummelkamp, T.R., Velds, A., Heimerikx, M., Kerkhoven, R.M., Madiredjo, M., Nijkamp, W., Weigelt, B., et al. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature *428*, 431-437. 10.1038/nature02371.
Beyond 1 Million Genomes (2021). 1+ Million Genomes Initiative. https://b1mg-project.eu/.
Bhinder, B., and Djaballah, H. (2013). Systematic analysis of RNAi reports identifies dismal commonality at gene-level and reveals an unprecedented enrichment in pooled shRNA screens. Comb Chem High Throughput Screen *16*, 665-681. 10.2174/13862073113169990045.
Bieging, K.T., Mello, S.S., and Attardi, L.D. (2014). Unravelling mechanisms of p53-mediated tumour suppression. Nat Rev Cancer *14*, 359-370. 10.1038/nrc3711.
Bird, A. (1992). The essentials of DNA methylation. Cell *70*, 5-8. 10.1016/0092-8674(92)90526-i.
Birmingham, A., Anderson, E.M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., et al. (2006). 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. Nat Methods *3*, 199-204. 10.1038/nmeth854.
Bitler, B.G., Aird, K.M., Garipov, A., Li, H., Amatangelo, M., Kossenkov, A.V., Schultz, D.C., Liu, Q., Shih Ie, M., Conejo-Garcia, J.R., et al. (2015). Synthetic lethality by targeting EZH2 methyltransferase activity in ARID1A-mutated cancers. Nat Med *21*, 231-238. 10.1038/nm.3799.
Blighe, K., and Lun, A. (2020). PCAtools: PCAtools: Everything Principal Components Analysis. https://github.com/kevinblighe/PCAtools.
Bridges, C.B. (1922). The origin of variations in sexual and sex-limited characters. The American Naturalist *56*, 51-63.
Brown, E.J., and Baltimore, D. (2003). Essential and dispensable roles of ATR in cell cycle arrest and genome maintenance. Genes Dev *17*, 615-628. 10.1101/gad.1067403.
Brown, J.S., O'Carrigan, B., Jackson, S.P., and Yap, T.A. (2017). Targeting DNA Repair in Cancer: Beyond PARP Inhibitors. Cancer Discov *7*, 20-37. 10.1158/2159-8290.CD-16-0860.
Brunen, D., and Bernards, R. (2017). Drug therapy: Exploiting synthetic lethality to improve cancer therapy. Nat Rev Clin Oncol *14*, 331-332. 10.1038/nrclinonc.2017.46.

Brunner, S.F., Roberts, N.D., Wylie, L.A., Moore, L., Aitken, S.J., Davies, S.E., Sanders, M.A., Ellis, P., Alder, C., Hooks, Y., et al. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature *574*, 538-542. 10.1038/s41586-019-1670-9.
Bunting, S.F., Callen, E., Wong, N., Chen, H.T., Polato, F., Gunn, A., Bothmer, A., Feldhahn, N., Fernandez-Capetillo, O., Cao, L., et al. (2010). 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. Cell *141*, 243-254. 10.1016/j.cell.2010.03.012.
Cairns, J. (1975). Mutation selection and the natural history of cancer. Nature *255*, 197-200. 10.1038/255197a0.
Caldas, C., and Venkitaraman, A.R. (2001). Tumor Suppressor Genes. In Encyclopedia of Genetics, S. Brenner, and J.H. Miller, eds. (Academic Press), pp. 2081-2088. https://doi.org/10.1006/rwgn.2001.1345.
Cancer Cell Line Encyclopedia, C., and Genomics of Drug Sensitivity in Cancer, C. (2015). Pharmacogenomic agreement between two cancer cell line data sets. Nature *528*, 84-87. 10.1038/nature15736.
Cancer Genome Atlas Research, N., Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A., Hoadley, K., Triche, T.J., Jr., Laird, P.W., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med *368*, 2059-2074. 10.1056/NEJMoa1301689.
Casciari, J.J., Sotirchos, S.V., and Sutherland, R.M. (1992). Variations in tumor cell growth rates and metabolism with oxygen concentration, glucose concentration, and extracellular pH. J Cell Physiol *151*, 386-394. 10.1002/jcp.1041510220.
Castellana, S., Mazza, T., Capocefalo, D., Genov, N., Biagini, T., Fusilli, C., Scholkmann, F., Relogio, A., Hogenesch, J.B., and Mazzoccoli, G. (2018). Systematic Analysis of Mouse Genome Reveals Distinct Evolutionary and Functional Properties Among Circadian and Ultradian Genes. Front Physiol *9*, 1178. 10.3389/fphys.2018.01178.
Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., Hubbard, T., Jostins, L., Maltby, N., and Mahon-Pearson, J. (2017). The National Genomics Research and Healthcare Kowledgebase. figshare.
Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov *2*, 401-404. 10.1158/2159-8290.CD-12-0095.
Chakrabarti, A.M., Henser-Brownhill, T., Monserrat, J., Poetsch, A.R., Luscombe, N.M., and Scaffidi, P. (2019). Target-Specific Precision of CRISPR-Mediated Genome Editing. Mol Cell *73*, 699-713 e696. 10.1016/j.molcel.2018.11.031.
Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol *2017*. 10.1200/PO.17.00011.
Chan, N., Pires, I.M., Bencokova, Z., Coackley, C., Luoto, K.R., Bhogal, N., Lakshman, M., Gottipati, P., Oliver, F.J., Helleday, T., et al. (2010). Contextual synthetic lethality of cancer cell kill based on the tumor microenvironment. Cancer Res *70*, 8045-8054. 10.1158/0008-5472.CAN-10-2352.
Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny.
Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res *45*, D369-D379. 10.1093/nar/gkw1102.
Chatterjee, N., and Walker, G.C. (2017). Mechanisms of DNA damage, repair, and mutagenesis. Environ Mol Mutagen *58*, 235-263. 10.1002/em.22087.
Chen, H., Zhang, Z., Jiang, S., Li, R., Li, W., Zhao, C., Hong, H., Huang, X., Li, H., and Bo, X. (2020). New insights on human essential genes based on integrated analysis and

the construction of the HEGIAP web-based platform. Brief Bioinform *21*, 1397-1410. 10.1093/bib/bbz072.
Chen, W.H., Lu, G., Chen, X., Zhao, X.M., and Bork, P. (2017). OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res *45*, D940-D944. 10.1093/nar/gkw1013.
Chen, W.H., Trachana, K., Lercher, M.J., and Bork, P. (2012). Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol Biol Evol *29*, 1703-1706. 10.1093/molbev/mss014.
Chipman, K.C., and Singh, A.K. (2009). Predicting genetic interactions with random walks on biological networks. BMC Bioinformatics *10*, 17. 10.1186/1471-2105-10-17.
Chou, C.H., Shrestha, S., Yang, C.D., Chang, N.W., Lin, Y.L., Liao, K.W., Huang, W.C., Sun, T.H., Tu, S.J., Lee, W.H., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res *46*, D296-D302. 10.1093/nar/gkx1067.
Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res *19*, 1553-1561. 10.1101/gr.092619.109.
ClinicalTrials (2022). A database of privately and publicly funded clinical studies conducted around the world. https://clinicaltrials.gov/ct2/home.
Cohen, I., Poreba, E., Kamieniarz, K., and Schneider, R. (2011). Histone modifiers in cancer: friends or foes? Genes Cancer *2*, 631-647. 10.1177/1947601911417176.
Colom, B., Herms, A., Hall, M.W.J., Dentro, S.C., King, C., Sood, R.K., Alcolea, M.P., Piedrafita, G., Fernandez-Antoran, D., Ong, S.H., et al. (2021). Mutant clones in normal epithelium outcompete and eliminate emerging tumours. Nature. 10.1038/s41586-021-03965-7.
Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585. 10.1038/ng.2653.
Consortium, G.T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318-1330. 10.1126/science.aaz1776.
Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*. 10.1126/science.aaf1420.
Csardi, G., and Nepusz, T. (2005). The Igraph Software Package for Complex Network Research. InterJournal *Complex Systems*, 1695.
Cusnir, M., and Cavalcante, L. (2012). Inter-tumor heterogeneity. Hum Vaccin Immunother *8*, 1143-1145. 10.4161/hv.21203.
D'Antonio, M., and Ciccarelli, F.D. (2011). Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol *7*, e1002029. 10.1371/journal.pcbi.1002029.
D'Antonio, M., Guerra, R.F., Cereda, M., Marchesi, S., Montani, F., Nicassio, F., Di Fiore, P.P., and Ciccarelli, F.D. (2013). Recessive cancer genes engage in negative genetic interactions with their functional paralogs. Cell Rep *5*, 1519-1526. 10.1016/j.celrep.2013.11.033.
D'Antonio, M., Pendino, V., Sinha, S., and Ciccarelli, F.D. (2012). Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. Nucleic Acids Res *40*, D978-983. 10.1093/nar/gkr952.
Daniels, D., Ford, M., Schwinn, M., Benink, H., Galbraith, M., Amunugama, R., Jones, R., Allen, D., Okazaki, N., Yamakawa, H., et al. (2013). Mutual exclusivity of MED12/MED12L, MED13/13L, and CDK8/19 paralogs revealed within the CDK-Mediator kinase module. Journal of Proteomics & Bioinformatics *S2*. 10.4172/jpb.S2-004.
Das, S., Camphausen, K., and Shankavaram, U. (2019). Pan-Cancer Analysis of Potential Synthetic Lethal Drug Targets Specific to Alterations in DNA Damage Response. Front Oncol *9*, 1136. 10.3389/fonc.2019.01136.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol *6*, e1001025. 10.1371/journal.pcbi.1001025.
De Kegel, B., Quinn, N., Thompson, N.A., Adams, D.J., and Ryan, C.J. (2021). Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines. Cell Systems. https://doi.org/10.1016/j.cels.2021.08.006.
De Kegel, B., and Ryan, C.J. (2019). Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. PLoS Genet *15*, e1008466. 10.1371/journal.pgen.1008466.
Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res *22*, 1589-1598. 10.1101/gr.134635.111.
Dempster, J.M., Rossen, J., Kazachkova, M., Pan, J., Kugener, G., Root, D.E., and Thsherniak, A. (2019). Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. . BioRxiv *720243*.
Dentro, S.C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., Yu, K., Rubanova, Y., Macintyre, G., Demeulemeester, J., et al. (2021). Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell *184*, 2239-2254 e2239. 10.1016/j.cell.2021.03.009.
DepMap Broad (2019a). DepMap GeCKO 19Q1. figshare. Fileset. doi:10.6084/m9.figshare.7668407.
DepMap Broad (2019b). Project SCORE processed with CERES. figshare. Dataset. doi:10.6084/m9.figshare.9116732.
DepMap Broad (2020). DepMap 20Q3 Public, figshare. Dataset doi:10.6084/m9.figshare.12931238.v1.
Dey, P., Baddour, J., Muller, F., Wu, C.C., Wang, H., Liao, W.-T., Lan, Z., Chen, A., Gutschner, T., Kang, Y., et al. (2017). Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. Nature *542*, 119-123. 10.1038/nature21052.
Dickerson, J.E., Zhu, A., Robertson, D.L., and Hentges, K.E. (2011). Defining the role of essential genes in human disease. PLoS One *6*, e27368. 10.1371/journal.pone.0027368.
Dittmer, D., Pati, S., Zambetti, G., Chu, S., Teresky, A.K., Moore, M., Finlay, C., and Levine, A.J. (1993). Gain of function mutations in p53. Nat Genet *4*, 42-46. 10.1038/ng0593-42.
Dobzhansky, T. (1946). Genetics of natural populations; recombination and variability in populations of Drosophila pseudoobscura. Genetics *31*, 269-290. 10.1093/genetics/31.3.269.
Dohner, H., Wei, A.H., and Lowenberg, B. (2021). Towards precision medicine for AML. Nat Rev Clin Oncol *18*, 577-590. 10.1038/s41571-021-00509-w.
Dolly, S.O., Gurden, M.D., Drosopoulos, K., Clarke, P., de Bono, J., Kaye, S., Workman, P., and Linardopoulos, S. (2017). RNAi screen reveals synthetic lethality between cyclin G-associated kinase and FBXW7 by inducing aberrant mitoses. Br J Cancer *117*, 954-964. 10.1038/bjc.2017.277.
Domazet-Loso, T., and Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol *25*, 2699-2707. 10.1093/molbev/msn214.
Domazet-Loso, T., and Tautz, D. (2010). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol *8*, 66. 10.1186/1741-7007-8-66.
Dressler, L., Bortolomeazzi, M., Keddar, M.R., Misetic, H., Sartini, G., Acha-Sagredo, A., Montorsi, L., Wijewardhane, N., Repana, D., Nulsen, J., et al. (2021). Comparative assessment of genes driving cancer and somatic evolution. bioRxiv, 2021.2008.2031.458177. 10.1101/2021.08.31.458177.

Dumbrava, E.I., and Meric-Bernstam, F. (2018). Personalized cancer therapy-leveraging a knowledge base for clinical decision-making. Cold Spring Harb Mol Case Stud *4*. 10.1101/mcs.a001578.
Egger, G., Liang, G., Aparicio, A., and Jones, P.A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. Nature *429*, 457-463. 10.1038/nature02625.
Elliott, K., and Larsson, E. (2021). Non-coding driver mutations in human cancer. Nat Rev Cancer *21*, 500-509. 10.1038/s41568-021-00371-z.
Ernst, P., Anastassiadis, K., Kranz, A., Stewart, A.F., Arndt, K., Waskow, C., Yokoyama, A., Jones, K., Neff, T., and Chen, Y. (2016). The Role of MLL1 and MLL2 in MLL Fusion Oncoprotein-Initiated Leukemia. Blood *128*, 573-573. 10.1182/blood.V128.22.573.573.
Esteller, M. (2006). Epigenetics provides a new generation of oncogenes and tumour-suppressor genes. Br J Cancer *94*, 179-183. 10.1038/sj.bjc.6602918.
Evers, B., Jastrzebski, K., Heijmans, J.P., Grernrum, W., Beijersbergen, R.L., and Bernards, R. (2016). CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. Nat Biotechnol *34*, 631-633. 10.1038/nbt.3536.
Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res *46*, D649-D655. 10.1093/nar/gkx1132.
Fang, H., Disteche, C.M., and Berletch, J.B. (2019). X Inactivation and Escape: Epigenetic and Structural Features. Front Cell Dev Biol *7*, 219. 10.3389/fcell.2019.00219.
Feinberg, A.P., Koldobskiy, M.A., and Gondor, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat Rev Genet *17*, 284-299. 10.1038/nrg.2016.13.
Feinberg, A.P., and Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature *301*, 89-92. 10.1038/301089a0.
Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A *105*, 4323-4328. 10.1073/pnas.0701722105.
Feng, Q., Wang, H., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Struhl, K., and Zhang, Y. (2002). Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. Curr Biol *12*, 1052-1058. 10.1016/s0960-9822(02)00901-6.
Ferlay J, L.M., Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2020). Global Cancer Observatory: Cancer Tomorrow. https://gco.iarc.fr/tomorrow.
Findlay, G.M. (2021). Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. Hum Mol Genet *30*, R187-R197. 10.1093/hmg/ddab219.
Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120-123. 10.1038/nature13695.
Fong, P.C., Boss, D.S., Yap, T.A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M.J., et al. (2009). Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med *361*, 123-134. 10.1056/NEJMoa0900212.
Fong, P.C., Yap, T.A., Boss, D.S., Carden, C.P., Mergui-Roelvink, M., Gourley, C., De Greve, J., Lubinski, J., Shanley, S., Messiou, C., et al. (2010). Poly(ADP)-ribose polymerase inhibition: frequent durable responses in BRCA carrier ovarian cancer correlating with platinum-free interval. J Clin Oncol *28*, 2512-2519. 10.1200/JCO.2009.26.9589.
Food and Drug Administration (2020). FDA approves tazemetostat for advanced epithelioid sarcoma. https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-tazemetostat-advanced-epithelioid-sarcoma.

Food and Drug Administration (2022). Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations. https://www.accessdata.fda.gov/scripts/cder/ob/index.cfm.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics *151*, 1531-1545. 10.1093/genetics/151.4.1531.
Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat Methods *11*, 801-807. 10.1038/nmeth.3027.
Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. Science *296*, 750-752. 10.1126/science.1068696.
Freilich, S., Massingham, T., Bhattacharyya, S., Ponsting, H., Lyons, P.A., Freeman, T.C., and Thornton, J.M. (2005). Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. Genome Biol *6*, R56. 10.1186/gb-2005-6-7-r56.
Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183. 10.1038/nrc1299.
Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics *25*, i54-62. 10.1093/bioinformatics/btp190.
Gauthier, M.L., Berman, H.K., Miller, C., Kozakeiwicz, K., Chew, K., Moore, D., Rabban, J., Chen, Y.Y., Kerlikowske, K., and Tlsty, T.D. (2007). Abrogated response to cellular stress identifies DCIS associated with subsequent tumor events and defines basal-like breast tumors. Cancer Cell *12*, 479-491. 10.1016/j.ccr.2007.10.017.
Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68-74. 10.1038/nature15393.
Gerlinger, M., Rowan, A.J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med *366*, 883-892. 10.1056/NEJMoa1113205.
Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., et al. (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122-128. 10.1038/s41586-019-1907-7.
Ghandi, M., Huang, F.W., Jane-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503-508. 10.1038/s41586-019-1186-3.
Gibney, E.R., and Nolan, C.M. (2010). Epigenetics and gene expression. Heredity (Edinb) *105*, 4-13. 10.1038/hdy.2010.54.
Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2019). CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res *47*, D559-D563. 10.1093/nar/gky973.
Godfrey, L., Crump, N.T., Thorne, R., Lau, I.J., Repapi, E., Dimou, D., Smith, A.L., Harman, J.R., Telenius, J.M., Oudelaar, A.M., et al. (2019). DOT1L inhibition reveals a distinct subset of enhancers dependent on H3K79 methylation. Nat Commun *10*, 2803. 10.1038/s41467-019-10844-3.
Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc Natl Acad Sci U S A *104*, 8685-8690. 10.1073/pnas.0701361104.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306-313. 10.1038/nature10762.
Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a Shared Vision for Cancer Genomic Data. N Engl J Med *375*, 1109-1112. 10.1056/NEJMp1607591.
Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. Nature *389*, 349-352. 10.1038/38664.
Guenther, M.G., Lane, W.S., Fischle, W., Verdin, E., Lazar, M.A., and Shiekhattar, R. (2000). A core SMRT corepressor complex containing HDAC3 and TBL1, a WD40-repeat protein linked to deafness. Genes Dev *14*, 1048-1057.
Gundle, K.R., Deutsch, G.B., Goodman, H.J., Pollack, S.M., Thompson, M.J., Davis, J.L., Lee, M.Y., Ramirez, D.C., Kerwin, W., Bertout, J.A., et al. (2020). Multiplexed Evaluation of Microdosed Antineoplastic Agents In Situ in the Tumor Microenvironment of Patients with Soft Tissue Sarcoma. Clin Cancer Res *26*, 3958-3968. 10.1158/1078-0432.CCR-20-0614.
Haber, J.E., Braberg, H., Wu, Q., Alexander, R., Haase, J., Ryan, C., Lipkin-Moore, Z., Franks-Skiba, K.E., Johnson, T., Shales, M., et al. (2013). Systematic triple-mutant analysis uncovers functional connectivity between pathways involved in chromosome regulation. Cell Rep *3*, 2168-2178. 10.1016/j.celrep.2013.05.007.
Han, M., Jia, L., Lv, W., Wang, L., and Cui, W. (2019). Epigenetic Enzyme Mutations: Role in Tumorigenesis and Molecular Inhibitors. Front Oncol *9*, 194. 10.3389/fonc.2019.00194.
Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70. 10.1016/s0092-8674(00)81683-9.
Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674. 10.1016/j.cell.2011.02.013.
Hargreaves, D.C., and Crabtree, G.R. (2011). ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. Cell Res *21*, 396-420. 10.1038/cr.2011.32.
Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R., and Moffat, J. (2014). Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol *10*, 733. 10.15252/msb.20145216.
Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515-1526. 10.1016/j.cell.2015.11.015.
Hart, T., and Moffat, J. (2016). BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics *17*, 164. 10.1186/s12859-016-1015-8.
Hartwell, L.H., Szankasi, P., Roberts, C.J., Murray, A.W., and Friend, S.H. (1997). Integrating genetic approaches into the discovery of anticancer drugs. Science *278*, 1064-1068. 10.1126/science.278.5340.1064.
Hase, T., Tanaka, H., Suzuki, Y., Nakagawa, S., and Kitano, H. (2009). Structure of protein interaction networks and their implications on drug design. PLoS Comput Biol *5*, e1000550. 10.1371/journal.pcbi.1000550.
Haynes, W.A., Tomczak, A., and Khatri, P. (2018). Gene annotation bias impedes biomedical research. Sci Rep *8*, 1362. 10.1038/s41598-018-19333-x.
He, N., Chan, C.K., Sobhian, B., Chou, S., Xue, Y., Liu, M., Alber, T., Benkirane, M., and Zhou, Q. (2011). Human Polymerase-Associated Factor complex (PAFc) connects the Super Elongation Complex (SEC) to RNA polymerase II on chromatin. Proc Natl Acad Sci U S A *108*, E636-645. 10.1073/pnas.1107107108.
Helming, K.C., Wang, X., Wilson, B.G., Vazquez, F., Haswell, J.R., Manchester, H.E., Kim, Y., Kryukov, G.V., Ghandi, M., Aguirre, A.J., et al. (2014). ARID1B is a specific vulnerability in ARID1A-mutant cancers. Nat Med *20*, 251-254. 10.1038/nm.3480.

Hoang, P.H., and Landi, M.T. (2022). DNA Methylation in Lung Cancer: Mechanisms and Associations with Histological Subtypes, Molecular Alterations, and Major Epidemiological Factors. Cancers (Basel) *14*. 10.3390/cancers14040961.
Hoffman, G.R., Rahal, R., Buxton, F., Xiang, K., McAllister, G., Frias, E., Bagdasarian, L., Huber, J., Lindeman, A., Chen, D., et al. (2014). Functional epigenetics approach identifies BRM/SMARCA2 as a critical synthetic lethal target in BRG1-deficient cancers. Proc Natl Acad Sci U S A *111*, 3128-3133. 10.1073/pnas.1316793111.
Honma, D., Kanno, O., Watanabe, J., Kinoshita, J., Hirasawa, M., Nosaka, E., Shiroishi, M., Takizawa, T., Yasumatsu, I., Horiuchi, T., et al. (2017). Novel orally bioavailable EZH1/2 dual inhibitors with greater antitumor efficacy than an EZH2 selective inhibitor. Cancer Sci *108*, 2069-2078. 10.1111/cas.13326.
Horiuchi, D., Camarda, R., Zhou, A.Y., Yau, C., Momcilovic, O., Balakrishnan, S., Corella, A.N., Eyob, H., Kessenbrock, K., Lawson, D.A., et al. (2016). PIM1 kinase inhibition as a targeted therapy against triple-negative breast tumors with elevated MYC expression. Nat Med *22*, 1321-1329. 10.1038/nm.4213.
Horowitz, L.F., Rodriguez, A.D., Dereli-Korkut, Z., Lin, R., Castro, K., Mikheev, A.M., Monnat, R.J., Jr., Folch, A., and Rostomily, R.C. (2020). Multiplexed drug testing of tumor slices using a microfluidic platform. NPJ Precis Oncol *4*, 12. 10.1038/s41698-020-0117-y.
Hsiau, T., Conant, D., Rossi, N., Maures, T., Waite, K., Yang, J., Joshi, S., Kelso, R., Holden, K., Enzmann, B., and Stoner, R. (2018). Inference of CRISPR Edits from Sanger Trace Data. bioRxiv.
Huang, H.Y., Lin, Y.C., Li, J., Huang, K.Y., Shrestha, S., Hong, H.C., Tang, Y., Chen, Y.G., Jin, C.N., Yu, Y., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res *48*, D148-D154. 10.1093/nar/gkz896.
Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res *44*, D286-293. 10.1093/nar/gkv1248.
Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res *47*, D309-D314. 10.1093/nar/gky1085.
Hughes, A.L., and Friedman, R. (2005). Gene duplication and the properties of biological networks. J Mol Evol *61*, 758-764. 10.1007/s00239-005-0037-z.
Huttlin, E.L., Bruckner, R.J., Navarrete-Perea, J., Cannon, J.R., Baltier, K., Gebreab, F., Gygi, M.P., Thornock, A., Zarraga, G., Tam, S., et al. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell *184*, 3022-3040 e3028. 10.1016/j.cell.2021.04.011.
Illingworth, R.S., and Bird, A.P. (2009). CpG islands--'a rough guide'. FEBS Lett *583*, 1713-1720. 10.1016/j.febslet.2009.04.012.
Ito, T., Young, M.J., Li, R., Jain, S., Wernitznig, A., Krill-Burger, J.M., Lemke, C.T., Monducci, D., Rodriguez, D.J., Chang, L., et al. (2021). Paralog knockout profiling identifies DUSP4 and DUSP6 as a digenic dependence in MAPK pathway-driven cancers. Nat Genet *53*, 1664-1672. 10.1038/s41588-021-00967-z.
Jackson, A.L., and Linsley, P.S. (2010). Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. Nat Rev Drug Discov *9*, 57-67. 10.1038/nrd3010.
Jaenisch, R., Beard, C., Lee, J., Marahrens, Y., and Panning, B. (1998). Mammalian X chromosome inactivation. Novartis Found Symp *214*, 200-209; discussion 209-213, 228-232. 10.1002/9780470515501.ch12.

Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet *33 Suppl*, 245-254. 10.1038/ng1089.
Jaspers, J.E., Kersbergen, A., Boon, U., Sol, W., van Deemter, L., Zander, S.A., Drost, R., Wientjens, E., Ji, J., Aly, A., et al. (2013). Loss of 53BP1 causes PARP inhibitor resistance in Brca1-mutated mouse mammary tumors. Cancer Discov *3*, 68-81. 10.1158/2159-8290.CD-12-0049.
Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res *48*, D498-D503. 10.1093/nar/gkz1031.
Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41-42. 10.1038/35075138.
Jerby-Arnon, L., Pfetzer, N., Waldman, Y.Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P.A., et al. (2014). Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell *158*, 1199-1209. 10.1016/j.cell.2014.07.027.
Jeziorska, D.M., Murray, R.J.S., De Gobbi, M., Gaentzsch, R., Garrick, D., Ayyub, H., Chen, T., Li, E., Telenius, J., Lynch, M., et al. (2017). DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. Proc Natl Acad Sci U S A *114*, E7526-E7535. 10.1073/pnas.1703087114.
Jia, D., Augert, A., Kim, D.W., Eastwood, E., Wu, N., Ibrahim, A.H., Kim, K.B., Dunn, C.T., Pillai, S.P.S., Gazdar, A.F., et al. (2018). Crebbp Loss Drives Small Cell Lung Cancer and Increases Sensitivity to HDAC Inhibition. Cancer Discov *8*, 1422-1437. 10.1158/2159-8290.CD-18-0385.
Jiang, Y., Ortega-Molina, A., Geng, H., Ying, H.Y., Hatzi, K., Parsa, S., McNally, D., Wang, L., Doane, A.S., Agirre, X., et al. (2017). CREBBP Inactivation Promotes the Development of HDAC3-Dependent Lymphomas. Cancer Discov *7*, 38-53. 10.1158/2159-8290.CD-16-0975.
Jimenez-Sanchez, G., Childs, B., and Valle, D. (2001). Human disease genes. Nature *409*, 853-855. 10.1038/35057050.
Jones, D.T., Jager, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y.J., Pugh, T.J., Hovestadt, V., Stutz, A.M., et al. (2012). Dissecting the genomic complexity underlying medulloblastoma. Nature *488*, 100-105. 10.1038/nature11284.
Jones, P.A., Issa, J.P., and Baylin, S. (2016). Targeting the cancer epigenome for therapy. Nat Rev Genet *17*, 630-641. 10.1038/nrg.2016.93.
Jonsson, P.F., and Bates, P.A. (2006). Global topological features of cancer proteins in the human interactome. Bioinformatics *22*, 2291-2297. 10.1093/bioinformatics/btl390.
Josse, J., and Husson, F. (2016). missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. Journal of Statistical Software *70*, 1-31. 0.18637/jss.v070.i01.
Kafri, R., Levy, M., and Pilpel, Y. (2006). The regulatory utilization of genetic redundancy through responsive backup circuits. Proc Natl Acad Sci U S A *103*, 11653-11658. 10.1073/pnas.0604883103.
Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J., and Kschischo, M. (2010). grofit: Fitting Biological Growth Curves with R. Journal of Statistical Software *33*, 1 - 21. 10.18637/jss.v033.i07.
Kakiuchi, N., and Ogawa, S. (2021). Clonal expansion in non-cancer tissues. Nat Rev Cancer *21*, 239-256. 10.1038/s41568-021-00335-3.
Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res *45*, D353-D361. 10.1093/nar/gkw1092.
Kaochar, S., Rusin, A., Foley, C., Rajapakshe, K., Robertson, M., Skapura, D., Mason, C., Berman De Ruiz, K., Tyryshkin, A.M., Deng, J., et al. (2021). Inhibition of GATA2 in

prostate cancer by a clinically available small molecule. Endocr Relat Cancer *29*, 15-31. 10.1530/ERC-21-0085.
Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443. 10.1038/s41586-020-2308-7.
Kaufman, B., Shapira-Frommer, R., Schmutzler, R.K., Audeh, M.W., Friedlander, M., Balmana, J., Mitchell, G., Fried, G., Stemmer, S.M., Hubert, A., et al. (2015). Olaparib monotherapy in patients with advanced cancer and a germline BRCA1/2 mutation. J Clin Oncol *33*, 244-250. 10.1200/JCO.2014.56.2728.
Kaur, J., Demokan, S., Tripathi, S.C., Macha, M.A., Begum, S., Califano, J.A., and Ralhan, R. (2010). Promoter hypermethylation in Indian primary oral squamous cell carcinoma. Int J Cancer *127*, 2367-2373. 10.1002/ijc.25377.
Kazanets, A., Shorstova, T., Hilmi, K., Marques, M., and Witcher, M. (2016). Epigenetic silencing of tumor suppressor genes: Paradigms, puzzles, and potential. Biochim Biophys Acta *1865*, 275-288. 10.1016/j.bbcan.2016.04.001.
Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664. 10.1101/gr.229202.
Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res *37*, D767-772. 10.1093/nar/gkn892.
Keshet, I., Lieman-Hurwitz, J., and Cedar, H. (1986). DNA methylation affects the formation of active chromatin. Cell *44*, 535-543. 10.1016/0092-8674(86)90263-1.
Klein, G., and Klein, E. (1985). Evolution of tumours and the impact of molecular oncology. Nature *315*, 190-195. 10.1038/315190a0.
Klijn, C., Durinck, S., Stawiski, E.W., Haverty, P.M., Jiang, Z., Liu, H., Degenhardt, J., Mayba, O., Gnad, F., Liu, J., et al. (2015). A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol *33*, 306-312. 10.1038/nbt.3080.
Klinghoffer, R.A., Bahrami, S.B., Hatton, B.A., Frazier, J.P., Moreno-Gonzalez, A., Strand, A.D., Kerwin, W.S., Casalini, J.R., Thirstrup, D.J., You, S., et al. (2015). A technology platform to assess multiple cancer agents simultaneously within a patient's tumor. Sci Transl Med *7*, 284ra258. 10.1126/scitranslmed.aaa7489.
Knutson, S.K., Warholic, N.M., Wigle, T.J., Klaus, C.R., Allain, C.J., Raimondi, A., Porter Scott, M., Chesworth, R., Moyer, M.P., Copeland, R.A., et al. (2013). Durable tumor regression in genetically altered malignant rhabdoid tumors by inhibition of methyltransferase EZH2. Proc Natl Acad Sci U S A *110*, 7922-7927. 10.1073/pnas.1303800110.
Ko, M., Huang, Y., Jankowska, A.M., Pape, U.J., Tahiliani, M., Bandukwala, H.S., An, J., Lamperti, E.D., Koh, K.P., Ganetzky, R., et al. (2010). Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature *468*, 839-843. 10.1038/nature09586.
Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693-705. 10.1016/j.cell.2007.02.005.
Krivtsov, A.V., and Armstrong, S.A. (2007). MLL translocations, histone modifications and leukaemia stem-cell development. Nat Rev Cancer *7*, 823-833. 10.1038/nrc2253.
Kroll, E.S., Hyland, K.M., Hieter, P., and Li, J.J. (1996). Establishing genetic interactions by a synthetic dosage lethality phenotype. Genetics *143*, 95-102. 10.1093/genetics/143.1.95.
Krzyszczyk, P., Acevedo, A., Davidoff, E.J., Timmins, L.M., Marrero-Berrios, I., Patel, M., White, C., Lowe, C., Sherba, J.J., Hartmanshenn, C., et al. (2018). The growing role of precision and personalized medicine for cancer treatment. Technology (Singap World Sci) *6*, 79-100. 10.1142/S2339547818300020.

Ku, A.A., Hu, H.M., Zhao, X., Shah, K.N., Kongara, S., Wu, D., McCormick, F., Balmain, A., and Bandyopadhyay, S. (2020). Integration of multiple biological contexts reveals principles of synthetic lethality that affect reproducibility. Nat Commun *11*, 2375. 10.1038/s41467-020-16078-y.
Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc *4*, 1073-1081. 10.1038/nprot.2009.86.
Kuuluvainen, E., Domenech-Moreno, E., Niemela, E.H., and Makela, T.P. (2018). Depletion of Mediator Kinase Module Subunits Represses Superenhancer-Associated Genes in Colon Cancer Cells. Mol Cell Biol *38*. 10.1128/MCB.00573-17.
Kwok, M., Davies, N., Agathanggelou, A., Smith, E., Oldreive, C., Petermann, E., Stewart, G., Brown, J., Lau, A., Pratt, G., et al. (2016). ATR inhibition induces synthetic lethality and overcomes chemoresistance in TP53- or ATM-defective chronic lymphocytic leukemia cells. Blood *127*, 582-595. 10.1182/blood-2015-05-644872.
Lac, V., Nazeran, T.M., Tessier-Cloutier, B., Aguirre-Hernandez, R., Albert, A., Lum, A., Khattra, J., Praetorius, T., Mason, M., Chiu, D., et al. (2019). Oncogenic mutations in histologically normal endometrium: the new normal? The Journal of Pathology *249*, 173-181. 10.1002/path.5314.
Lac, V., Verhoef, L., Aguirre-Hernandez, R., Nazeran, T.M., Tessier-Cloutier, B., Praetorius, T., Orr, N.L., Noga, H., Lum, A., Khattra, J., et al. (2018). Iatrogenic endometriosis harbors somatic cancer-driver mutations. Human Reproduction *34*, 69-78. 10.1093/humrep/dey332.
Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218. 10.1038/nature12213.
Lawson, A.R.J., Abascal, F., Coorens, T.H.H., Hooks, Y., O'Neill, L., Latimer, C., Raine, K., Sanders, M.A., Warren, A.Y., Mahbubani, K.T.A., et al. (2020). Extensive heterogeneity in somatic mutation and selection in the human bladder. Science *370*, 75-82. 10.1126/science.aba8347.
Lee, J.S., Das, A., Jerby-Arnon, L., Arafeh, R., Auslander, N., Davidson, M., McGarry, L., James, D., Amzallag, A., Park, S.G., et al. (2018). Harnessing synthetic lethality to predict the response to cancer treatment. Nat Commun *9*, 2546. 10.1038/s41467-018-04647-1.
Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., et al. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. Nature *574*, 532-537. 10.1038/s41586-019-1672-7.
Lenoir, W.F., Lim, T.L., and Hart, T. (2018). PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. Nucleic Acids Res *46*, D776-D780. 10.1093/nar/gkx993.
Letai, A. (2017). Functional precision cancer medicine-moving beyond pure genomics. Nat Med *23*, 1028-1035. 10.1038/nm.4389.
Lever, J., Zhao, E.Y., Grewal, J., Jones, M.R., and Jones, S.J.M. (2019). CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods *16*, 505-507. 10.1038/s41592-019-0422-y.
Ley, T.J., Ding, L., Walter, M.J., McLellan, M.D., Lamprecht, T., Larson, D.E., Kandoth, C., Payton, J.E., Baty, J., Welch, J., et al. (2010). DNMT3A mutations in acute myeloid leukemia. N Engl J Med *363*, 2424-2433. 10.1056/NEJMoa1005143.
Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66-72. 10.1038/nature07485.

Li, J., Wang, J., Wang, J., Nawaz, Z., Liu, J.M., Qin, J., and Wong, J. (2000). Both corepressor proteins SMRT and N-CoR exist in large protein complexes containing HDAC3. EMBO J *19*, 4342-4350. 10.1093/emboj/19.16.4342.
Li, S., Topatana, W., Juengpanich, S., Cao, J., Hu, J., Zhang, B., Ma, D., Cai, X., and Chen, M. (2020). Development of synthetic lethality in cancer: molecular and cellular classification. Signal Transduct Target Ther *5*, 241. 10.1038/s41392-020-00358-6.
Li, Y., Wen, H., Xi, Y., Tanaka, K., Wang, H., Peng, D., Ren, Y., Jin, Q., Dent, S.Y., Li, W., et al. (2014). AF9 YEATS domain links histone acetylation to DOT1L-mediated H3K79 methylation. Cell *159*, 558-571. 10.1016/j.cell.2014.09.049.
Li, Y.H., and Zhang, G.G. (2017). Network-based characterization and prediction of human DNA repair genes and pathways. Sci Rep *8*, 45714. 10.1038/srep45714.
Liang, H., and Li, W.H. (2007a). Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet *23*, 375-378. 10.1016/j.tig.2007.04.005.
Liang, H., and Li, W.H. (2007b). MicroRNA regulation of human protein protein interaction network. RNA *13*, 1402-1408. 10.1261/rna.634607.
Liao, B.Y., and Zhang, J. (2007). Mouse duplicate genes are as essential as singletons. Trends Genet *23*, 378-381. 10.1016/j.tig.2007.05.006.
Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18-22.
Liu, E.M., Martinez-Fundichely, A., Bollapragada, R., Spiewack, M., and Khurana, E. (2021). CNCDatabase: a database of non-coding cancer drivers. Nucleic Acids Res *49*, D1094-D1101. 10.1093/nar/gkaa915.
Liu, L., Ulbrich, J., Muller, J., Wustefeld, T., Aeberhard, L., Kress, T.R., Muthalagu, N., Rycak, L., Rudalska, R., Moll, R., et al. (2012). Deregulated MYC expression induces dependence upon AMPK-related kinase 5. Nature *483*, 608-612. 10.1038/nature10927.
Liu, S.H., Shen, P.C., Chen, C.Y., Hsu, A.N., Cho, Y.C., Lai, Y.L., Chen, F.H., Li, C.Y., Wang, S.C., Chen, M., et al. (2020). DriverDBv3: a multi-omics database for cancer driver gene research. Nucleic Acids Res *48*, D863-D870. 10.1093/nar/gkz964.
Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat *37*, 235-241. 10.1002/humu.22932.
Liu, Y., Sun, J., and Zhao, M. (2017). ONGene: A literature-based database for human oncogenes. J Genet Genomics *44*, 119-121. 10.1016/j.jgg.2016.12.004.
Lodish, H., Berk, A., Zipursky, S., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). Molecular Cell Biology, 4th Edition (W. H. Freeman).
Lord, C.J., and Ashworth, A. (2017). PARP inhibitors: Synthetic lethality in the clinic. Science *355*, 1152-1158. 10.1126/science.aam7344.
Lord, C.J., Quinn, N., and Ryan, C.J. (2020). Integrative analysis of large-scale loss-of-function screens identifies robust cancer-associated genetic interactions. Elife *9*. 10.7554/eLife.58925.
Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.K., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell *137*, 835-848. 10.1016/j.cell.2009.05.006.
Magen, A., Das Sahu, A., Lee, J.S., Sharmin, M., Lugo, A., Gutkind, J.S., Schaffer, A.A., Ruppin, E., and Hannenhalli, S. (2019). Beyond Synthetic Lethality: Charting the Landscape of Pairwise Gene Expression States Associated with Survival in Cancer. Cell Rep *28*, 938-948 e936. 10.1016/j.celrep.2019.06.067.
Makino, T., Hokamp, K., and McLysaght, A. (2009). The complex relationship of gene duplication and essentiality. Trends Genet *25*, 152-155. 10.1016/j.tig.2009.03.001.
Malone, C.F., Dharia, N.V., Kugener, G., Forman, A.B., Rothberg, M.V., Abdusamad, M., Gonzalez, A., Kuljanin, M., Robichaud, A.L., Conway, A.S., et al. (2021). Selective Modulation of a Pan-Essential Protein as a Therapeutic Strategy in Cancer. Cancer Discov *11*, 2282-2299. 10.1158/2159-8290.CD-20-1213.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L.Y., et al. (2012). Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov *2*, 172-189. 10.1158/2159-8290.CD-11-0224.

Martincorena, I. (2019). Somatic mutation and clonal expansions in human tissues. Genome Med *11*, 35. 10.1186/s13073-019-0648-4.

Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science *349*, 1483-1489. 10.1126/science.aab4082.

Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F., Hall, M.W.J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M.R., et al. (2018). Somatic mutant clones colonize the human esophagus with age. Science *362*, 911-917. 10.1126/science.aau3879.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029-1041 e1021. 10.1016/j.cell.2017.09.042.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science *348*, 880-886. 10.1126/science.aaa6806.

Martinez-Jimenez, F., Muinos, F., Sentis, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., et al. (2020). A compendium of mutational cancer driver genes. Nat Rev Cancer *20*, 555-572. 10.1038/s41568-020-0290-x.

Mashtalir, N., D'Avino, A.R., Michel, B.C., Luo, J., Pan, J., Otto, J.E., Zullow, H.J., McKenzie, Z.M., Kubiak, R.L., St Pierre, R., et al. (2018). Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. Cell *175*, 1272-1288 e1220. 10.1016/j.cell.2018.09.032.

McDonald, E.R., 3rd, de Weck, A., Schlabach, M.R., Billy, E., Mavrakis, K.J., Hoffman, G.R., Belur, D., Castelletti, D., Frias, E., Gampa, K., et al. (2017). Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. Cell *170*, 577-592 e510. 10.1016/j.cell.2017.07.005.

McFarland, J.M., Ho, Z.V., Kugener, G., Dempster, J.M., Montgomery, P.G., Bryan, J.G., Krill-Burger, J.M., Green, T.M., Vazquez, F., Boehm, J.S., et al. (2018). Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat Commun *9*, 4610. 10.1038/s41467-018-06916-5.

McGranahan, N., and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell *168*, 613-628. 10.1016/j.cell.2017.01.018.

McKusick-Nathans Institute of Genetic Medicine (2020). Online Mendelian Inheritance in Man, OMIM.

McLornan, D.P., List, A., and Mufti, G.J. (2014). Applying synthetic lethality for the selective targeting of cancer. N Engl J Med *371*, 1725-1735. 10.1056/NEJMra1407390.

Medvedeva, Y.A., Lennartsson, A., Ehsani, R., Kulakovskiy, I.V., Vorontsov, I.E., Panahandeh, P., Khimulya, G., Kasukawa, T., Consortium, F., and Drablos, F. (2015). EpiFactors: a comprehensive database of human epigenetic factors and complexes. Database (Oxford) *2015*, bav067. 10.1093/database/bav067.

Meijer, T.G., Naipal, K.A., Jager, A., and van Gent, D.C. (2017). Ex vivo tumor culture systems for functional drug testing and therapy response prediction. Future Sci OA *3*, FSO190. 10.4155/fsoa-2017-0003.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B., et al. (2012). Systematic dissection and

optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol *30*, 271-277. 10.1038/nbt.2137.
Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet *49*, 1779-1784. 10.1038/ng.3984.
Moore, L., Leongamornlert, D., Coorens, T.H.H., Sanders, M.A., Ellis, P., Dentro, S.C., Dawson, K.J., Butler, T., Rahbari, R., Mitchell, T.J., et al. (2020). The mutational landscape of normal human endometrial epithelium. Nature *580*, 640-646. 10.1038/s41586-020-2214-z.
Morris, L.G., and Chan, T.A. (2015). Therapeutic targeting of tumor suppressor genes. Cancer *121*, 1357-1368. 10.1002/cncr.29140.
Mourikis, T.P., Benedetti, L., Foxall, E., Temelkovski, D., Nulsen, J., Perner, J., Cereda, M., Lagergren, J., Howell, M., Yau, C., et al. (2019). Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. Nat Commun *10*, 3101. 10.1038/s41467-019-10898-3.
Munoz, D.M., Cassiani, P.J., Li, L., Billy, E., Korn, J.M., Jones, M.D., Golji, J., Ruddy, D.A., Yu, K., McAllister, G., et al. (2016). CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. Cancer Discov *6*, 900-913. 10.1158/2159-8290.CD-16-0178.
Nakamura, Y., Kawazoe, A., Lordick, F., Janjigian, Y.Y., and Shitara, K. (2021). Biomarker-targeted therapies for advanced-stage gastric and gastro-oesophageal junction cancers: an emerging paradigm. Nat Rev Clin Oncol *18*, 473-487. 10.1038/s41571-021-00492-2.
Nakata, Y., Brignier, A.C., Jin, S., Shen, Y., Rudnick, S.I., Sugita, M., and Gewirtz, A.M. (2010). c-Myb, Menin, GATA-3, and MLL form a dynamic transcription complex that plays a pivotal role in human T helper type 2 cell development. Blood *116*, 1280-1290. 10.1182/blood-2009-05-223255.
Neggers, J.E., Paolella, B.R., Asfaw, A., Rothberg, M.V., Skipper, T.A., Yang, A., Kalekar, R.L., Krill-Burger, J.M., Dharia, N.V., Kugener, G., et al. (2020). Synthetic Lethal Interaction between the ESCRT Paralog Enzymes VPS4A and VPS4B in Cancers Harboring Loss of Chromosome 18q or 16q. Cell Rep *33*, 108493. 10.1016/j.celrep.2020.108493.
Newell-Price, J., Clark, A.J., and King, P. (2000). DNA methylation and silencing of gene expression. Trends Endocrinol Metab *11*, 142-148. 10.1016/s1043-2760(00)00248-4.
Nijman, S.M. (2011). Synthetic lethality: general principles, utility and detection using genetic screens in human cells. FEBS Lett *585*, 1-6. 10.1016/j.febslet.2010.11.024.
Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23-28. 10.1126/science.959840.
Nulsen, J., Misetic, H., Yau, C., and Ciccarelli, F.D. (2021). Pan-cancer detection of driver genes at the single-patient resolution. Genome Med *13*, 12. 10.1186/s13073-021-00830-0.
O'Brien, S.G., Guilhot, F., Larson, R.A., Gathmann, I., Baccarani, M., Cervantes, F., Cornelissen, J.J., Fischer, T., Hochhaus, A., Hughes, T., et al. (2003). Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. N Engl J Med *348*, 994-1004. 10.1056/NEJMoa022457.
O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res *44*, D733-745. 10.1093/nar/gkv1189.
O'Neil, N.J., Bailey, M.L., and Hieter, P. (2017). Synthetic lethality and cancer. Nat Rev Genet *18*, 613-623. 10.1038/nrg.2017.47.

Ogiwara, H., Sasaki, M., Mitachi, T., Oike, T., Higuchi, S., Tominaga, Y., and Kohno, T. (2016). Targeting p300 Addiction in CBP-Deficient Cancers Causes Synthetic Lethality by Apoptotic Cell Death due to Abrogation of MYC Expression. Cancer Discov *6*, 430-445. 10.1158/2159-8290.CD-15-0754.
Oike, T., Ogiwara, H., Tominaga, Y., Ito, K., Ando, O., Tsuta, K., Mizukami, T., Shimada, Y., Isomura, H., Komachi, M., et al. (2013). A synthetic lethality-based strategy to treat cancers harboring a genetic deficiency in the chromatin remodeling factor BRG1. Cancer Res *73*, 5508-5518. 10.1158/0008-5472.CAN-12-4593.
Olafsson, S., McIntyre, R.E., Coorens, T., Butler, T., Jung, H., Robinson, P.S., Lee-Six, H., Sanders, M.A., Arestang, K., Dawson, C., et al. (2020). Somatic Evolution in Non-neoplastic IBD-Affected Colon. Cell. 10.1016/j.cell.2020.06.036.
Ooms, J., James, D., DebRoy, S., Wickham, H., and Horner, J. (2021). RMySQL: Database Interface and 'MySQL' Driver for R. R package version 0.10.22. https://CRAN.R-project.org/package=RMySQL.
Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res *42*, D358-363. 10.1093/nar/gkt1115.
Ostrow, S.L., Barshir, R., DeGregori, J., Yeger-Lotem, E., and Hershberg, R. (2014). Cancer evolution is associated with pervasive positive selection on globally expressed genes. PLoS Genet *10*, e1004239. 10.1371/journal.pgen.1004239.
Oughtred, R., Stark, C., Breitkreutz, B.J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., et al. (2019). The BioGRID interaction database: 2019 update. Nucleic Acids Res *47*, D529-D541. 10.1093/nar/gky1079.
Owen-Hughes, T., Utley, R.T., Cote, J., Peterson, C.L., and Workman, J.L. (1996). Persistent site-specific remodeling of a nucleosome array by transient action of the SWI/SNF complex. Science *273*, 513-516. 10.1126/science.273.5274.513.
Paladugu, S.R., Zhao, S., Ray, A., and Raval, A. (2008). Mining protein networks for synthetic genetic interactions. BMC Bioinformatics *9*, 426. 10.1186/1471-2105-9-426.
Papp, B., Pal, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194-197. 10.1038/nature01771.
Parrish, P.C.R., Thomas, J.D., Gabel, A.M., Kamlapurkar, S., Bradley, R.K., and Berger, A.H. (2021). Discovery of synthetic lethal and tumor suppressor paralog pairs in the human genome. Cell Rep *36*, 109597. 10.1016/j.celrep.2021.109597.
Parsons, D.W., Li, M., Zhang, X., Jones, S., Leary, R.J., Lin, J.C., Boca, S.M., Carter, H., Samayoa, J., Bettegowda, C., et al. (2011). The genetic landscape of the childhood cancer medulloblastoma. Science *331*, 435-439. 10.1126/science.1198056.
Pathak, H.B., Zhou, Y., Sethi, G., Hirst, J., Schilder, R.J., Golemis, E.A., and Godwin, A.K. (2015). A Synthetic Lethality Screen Using a Focused siRNA Library to Identify Sensitizers to Dasatinib Therapy for the Treatment of Epithelial Ovarian Cancer. PLoS One *10*, e0144126. 10.1371/journal.pone.0144126.
Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol *30*, 265-270. 10.1038/nbt.2136.
Perrier, V., Meyer, F., and Granjon, D. (2021). shinyWidgets: Custom Inputs Widgets for Shiny. R package version 0.6.1. https://CRAN.R-project.org/package=shinyWidgets.
Pfister, S.X., Markkanen, E., Jiang, Y., Sarkar, S., Woodcock, M., Orlando, G., Mavrommati, I., Pai, C.C., Zalmas, L.P., Drobnitzky, N., et al. (2015). Inhibiting WEE1 Selectively Kills Histone H3K36me3-Deficient Cancers by dNTP Starvation. Cancer Cell *28*, 557-568. 10.1016/j.ccell.2015.09.015.

Plass, C., Pfister, S.M., Lindroth, A.M., Bogatyrova, O., Claus, R., and Lichter, P. (2013). Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. Nat Rev Genet *14*, 765-780. 10.1038/nrg3554.
Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res *20*, 110-121. 10.1101/gr.097857.109.
Prachumwat, A., and Li, W.H. (2006). Protein function, connectivity, and duplicability in yeast. Mol Biol Evol *23*, 30-39. 10.1093/molbev/msi249.
Pugh, T.J., Weeraratne, S.D., Archer, T.C., Pomeranz Krummel, D.A., Auclair, D., Bochicchio, J., Carneiro, M.O., Carter, S.L., Cibulskis, K., Erlich, R.L., et al. (2012). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature *488*, 106-110. 10.1038/nature11329.
R Core Team (2020). R: A Language and Environment for Statistical Computing. https://www.R-project.org/.
Rambaldi, D., Giorgi, F.M., Capuani, F., Ciliberto, A., and Ciccarelli, F.D. (2008). Low duplicability and network fragility of cancer genes. Trends Genet *24*, 427-430. 10.1016/j.tig.2008.06.003.
Ramon, Y.C.S., Sese, M., Capdevila, C., Aasen, T., De Mattos-Arruda, L., Diaz-Cano, S.J., Hernandez-Losa, J., and Castellvi, J. (2020). Clinical implications of intratumor heterogeneity: challenges and opportunities. J Mol Med (Berl) *98*, 161-177. 10.1007/s00109-020-01874-2.
Rampias, T., Karagiannis, D., Avgeris, M., Polyzos, A., Kokkalis, A., Kanaki, Z., Kousidou, E., Tzetis, M., Kanavakis, E., Stravodimos, K., et al. (2019). The lysine-specific methyltransferase KMT2C/MLL3 regulates DNA repair components in cancer. EMBO Rep *20*. 10.15252/embr.201846821.
Rehman, F.L., Lord, C.J., and Ashworth, A. (2010). Synthetic lethal approaches to breast cancer therapy. Nat Rev Clin Oncol *7*, 718-724. 10.1038/nrclinonc.2010.172.
Reid, R.J., Du, X., Sunjevaric, I., Rayannavar, V., Dittmar, J., Bryant, E., Maurer, M., and Rothstein, R. (2016). A Synthetic Dosage Lethal Genetic Interaction Between CKS1B and PLK1 Is Conserved in Yeast and Human Cancer Cells. Genetics *204*, 807-819. 10.1534/genetics.116.190231.
Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S.K., Tourna, A., Yakovleva, A., Palmieri, T., and Ciccarelli, F.D. (2019). The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol *20*, 1. 10.1186/s13059-018-1612-0.
Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res *39*, e118. 10.1093/nar/gkr407.
Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res *38*, D497-501. 10.1093/nar/gkp914.
Ruzankina, Y., Schoppy, D.W., Asare, A., Clark, C.E., Vonderheide, R.H., and Brown, E.J. (2009). Tissue regenerative delays and synthetic lethality in adult mice after combined deletion of Atr and Trp53. Nat Genet *41*, 1144-1149. 10.1038/ng.441.
Saha, A., Wittmeyer, J., and Cairns, B.R. (2006). Chromatin remodelling: the industrial revolution of DNA around histones. Nat Rev Mol Cell Biol *7*, 437-447. 10.1038/nrm1945.
Saito, Y., Koya, J., Araki, M., Kogure, Y., Shingaki, S., Tabata, M., McClure, M.B., Yoshifuji, K., Matsumoto, S., Isaka, Y., et al. (2020). Landscape and function of multiple mutations within individual oncogenes. Nature *582*, 95-99. 10.1038/s41586-020-2175-2.
Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res *32*, D449-451. 10.1093/nar/gkh086.

Sato, S., Tomomori-Sato, C., Parmely, T.J., Florens, L., Zybailov, B., Swanson, S.K., Banks, C.A., Jin, J., Cai, Y., Washburn, M.P., et al. (2004). A set of consensus mammalian mediator subunits identified by multidimensional protein identification technology. Mol Cell *14*, 685-691. 10.1016/j.molcel.2004.05.006.
Sauka-Spengler, T., Meulemans, D., Jones, M., and Bronner-Fraser, M. (2007). Ancient evolutionary origin of the neural crest gene regulatory network. Dev Cell *13*, 405-420. 10.1016/j.devcel.2007.08.005.
Schulz, W.A., Steinhoff, C., and Florl, A.R. (2006). Methylation of endogenous human retroelements in health and disease. Curr Top Microbiol Immunol *310*, 211-250. 10.1007/3-540-31181-5_11.
Schwarz, J.M., Rodelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods *7*, 575-576. 10.1038/nmeth0810-575.
Shen, C., Jo, S.Y., Liao, C., Hess, J.L., and Nikolovska-Coleska, Z. (2013). Targeting recruitment of disruptor of telomeric silencing 1-like (DOT1L): characterizing the interactions between DOT1L and mixed lineage leukemia (MLL) fusion proteins. J Biol Chem *288*, 30585-30596. 10.1074/jbc.M113.457135.
Shen, J.P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A.N., et al. (2017). Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. Nat Methods *14*, 573-576. 10.1038/nmeth.4225.
Shen, L., Shi, Q., and Wang, W. (2018). Double agents: genes with both oncogenic and tumor-suppressor functions. Oncogenesis *7*, 25. 10.1038/s41389-018-0034-x.
Shendure, J., and Akey, J.M. (2015). The origins, determinants, and consequences of human mutations. Science *349*, 1478-1483. 10.1126/science.aaa9119.
Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat *34*, 57-65. 10.1002/humu.22225.
Shin, S.H., Bode, A.M., and Dong, Z. (2017). Precision medicine: the foundation of future cancer therapeutics. NPJ Precis Oncol *1*, 12. 10.1038/s41698-017-0016-z.
Siegfried, Z., and Cedar, H. (1997). DNA methylation: a molecular lock. Curr Biol *7*, R305-307. 10.1016/s0960-9822(06)00144-8.
Sinha, S., Thomas, D., Chan, S., Gao, Y., Brunen, D., Torabi, D., Reinisch, A., Hernandez, D., Chan, A., Rankin, E.B., et al. (2017). Systematic discovery of mutation-specific synthetic lethals by mining pan-cancer human primary tumor data. Nat Commun *8*, 15580. 10.1038/ncomms15580.
Sleutels, F., Soochit, W., Bartkuhn, M., Heath, H., Dienstbach, S., Bergmaier, P., Franke, V., Rosa-Garrido, M., van de Nobelen, S., Caesar, L., et al. (2012). The male germ cell gene regulator CTCFL is functionally different from CTCF and binds CTCF-like consensus sites in a nucleosome composition-dependent manner. Epigenetics Chromatin *5*, 8. 10.1186/1756-8935-5-8.
Smith, I., Greenside, P.G., Natoli, T., Lahr, D.L., Wadden, D., Tirosh, I., Narayan, R., Root, D.E., Golub, T.R., Subramanian, A., and Doench, J.G. (2017). Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. PLoS Biol *15*, e2003213. 10.1371/journal.pbio.2003213.
Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat Rev Genet *14*, 204-220. 10.1038/nrg3354.
Soussi, T., and Wiman, K.G. (2015). TP53: an oncogene in disguise. Cell Death Differ *22*, 1239-1249. 10.1038/cdd.2015.53.
Soutourina, J. (2018). Transcription regulation by the Mediator complex. Nat Rev Mol Cell Biol *19*, 262-274. 10.1038/nrm.2017.115.

Srivas, R., Shen, J.P., Yang, C.C., Sun, S.M., Li, J., Gross, A.M., Jensen, J., Licon, K., Bojorquez-Gomez, A., Klepper, K., et al. (2016). A Network of Conserved Synthetic Lethal Interactions for Exploration of Precision Cancer Therapy. Mol Cell *63*, 514-525. 10.1016/j.molcel.2016.06.022.
Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719-724. 10.1038/nature07943.
Strese, S., Fryknas, M., Larsson, R., and Gullbo, J. (2013). Effects of hypoxia on human cancer cell line chemosensitivity. BMC Cancer *13*, 331. 10.1186/1471-2407-13-331.
Suda, K., Nakaoka, H., Yoshihara, K., Ishiguro, T., Tamura, R., Mori, Y., Yamawaki, K., Adachi, S., Takahashi, T., Kase, H., et al. (2018). Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. Cell Rep *24*, 1777-1789. 10.1016/j.celrep.2018.07.037.
Sun, W., and Yang, J. (2010). Functional mechanisms for human tumor suppressors. J Cancer *1*, 136-140. 10.7150/jca.1.136.
Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin *71*, 209-249. 10.3322/caac.21660.
Syed, A.S., D'Antonio, M., and Ciccarelli, F.D. (2010). Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. Nucleic Acids Res *38*, D670-675. 10.1093/nar/gkp957.
Szczurek, E., Misra, N., and Vingron, M. (2013). Synthetic sickness or lethality points at candidate combination therapy targets in glioblastoma. Int J Cancer *133*, 2123-2132. 10.1002/ijc.28235.
Szedlak, A., Smith, N., Liu, L., Paternostro, G., and Piermarocchi, C. (2016). Evolutionary and Topological Properties of Genes and Community Structures in Human Gene Regulatory Networks. PLoS Comput Biol *12*, e1005009. 10.1371/journal.pcbi.1005009.
Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics *29*, 2238-2244. 10.1093/bioinformatics/btt395.
Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M.P., Vivancos, A., Rovira, A., Tusquets, I., Albanell, J., Rodon, J., Tabernero, J., et al. (2018). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med *10*, 25. 10.1186/s13073-018-0531-8.
Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res *47*, D941-D947. 10.1093/nar/gky1015.
TCGA Research Network (2021). https://www.cancer.gov/tcga. https://www.cancer.gov/tcga.
Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science *294*, 2364-2368. 10.1126/science.1065810.
Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., et al. (2004). Global mapping of the yeast genetic interaction network. Science *303*, 808-813. 10.1126/science.1091317.
Tremblay, M., Sanchez-Ferras, O., and Bouchard, M. (2018). GATA transcription factors in development and disease. Development *145*. 10.1242/dev.164384.
Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell *170*, 564-576 e516. 10.1016/j.cell.2017.06.010.

Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419. 10.1126/science.1260419.
van der Lelij, P., Lieb, S., Jude, J., Wutz, G., Santos, C.P., Falkenberg, K., Schlattl, A., Ban, J., Schwentner, R., Hoffmann, T., et al. (2017). Synthetic lethality between the cohesin subunits STAG1 and STAG2 in diverse cancer contexts. Elife *6*. 10.7554/eLife.26980.
Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., et al. (2010). Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A *107*, 16910-16915. 10.1073/pnas.1009843107.
Varga-Weisz, P.D., Wilm, M., Bonte, E., Dumas, K., Mann, M., and Becker, P.B. (1997). Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. Nature *388*, 598-602. 10.1038/41587.
Veitia, R.A. (2002). Exploring the etiology of haploinsufficiency. Bioessays *24*, 175-184. 10.1002/bies.10023.
Veitia, R.A. (2004). Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics *168*, 569-574. 10.1534/genetics.104.029785.
Vichas, A., Riley, A.K., Nkinsi, N.T., Kamlapurkar, S., Parrish, P.C.R., Lo, A., Duke, F., Chen, J., Fung, I., Watson, J., et al. (2021). Integrative oncogene-dependency mapping identifies RIT1 vulnerabilities and synergies in lung cancer. Nat Commun *12*, 4789. 10.1038/s41467-021-24841-y.
Viswanathan, S.R., Nogueira, M.F., Buss, C.G., Krill-Burger, J.M., Wawer, M.J., Malolepsza, E., Berger, A.C., Choi, P.S., Shih, J., Taylor, A.M., et al. (2018). Genome-scale analysis identifies paralog lethality as a vulnerability of chromosome 1p loss in cancer. Nat Genet *50*, 937-943. 10.1038/s41588-018-0155-3.
Vogel, C.L., Cobleigh, M.A., Tripathy, D., Gutheil, J.C., Harris, L.N., Fehrenbacher, L., Slamon, D.J., Murphy, M., Novotny, W.F., Burchmore, M., et al. (2002). Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol *20*, 719-726. 10.1200/JCO.2002.20.3.719.
Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558. 10.1126/science.1235122.
Vu, V.Q. (2011). ggbiplot: A ggplot2 based biplot. http://github.com/vqv/ggbiplot.
Wan, L., Wen, H., Li, Y., Lyu, J., Xi, Y., Hoshii, T., Joseph, J.K., Wang, X., Loh, Y.E., Erb, M.A., et al. (2017). ENL links histone acetylation to oncogenic gene expression in acute myeloid leukaemia. Nature *543*, 265-269. 10.1038/nature21687.
Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*, e164. 10.1093/nar/gkq603.
Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096-1101. 10.1126/science.aac7041.
Wang, W., Wei, Z., Lam, T.W., and Wang, J. (2011). Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci Rep *1*, 55. 10.1038/srep00055.
Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet *46*, 1160-1165. 10.1038/ng.3101.
Wickham, H., Müller, K., and R Special Interest Group on Databases (R-SIG-DB) (2021). DBI: R Database Interface. R package version 1.1.1. https://CRAN.R-project.org/package=DBI.

Wijewardhane, N., Dressler, L., and Ciccarelli, F.D. (2020). Normal Somatic Mutations in Cancer Transformation. Cancer Cell. 10.1016/j.ccell.2020.11.002.
Wilson, B.G., and Roberts, C.W. (2011). SWI/SNF nucleosome remodellers and cancer. Nat Rev Cancer *11*, 481-492. 10.1038/nrc3068.
World Health Organization (2020). Global Health Estimates 2019: Estimated deaths by age, sex and cause.
Wrzeszczynski, K.O., Varadan, V., Byrnes, J., Lum, E., Kamalakaran, S., Levine, D.A., Dimitrova, N., Zhang, M.Q., and Lucito, R. (2011). Identification of tumor suppressors and oncogenes from genomic and epigenetic features in ovarian cancer. PLoS One *6*, e28503. 10.1371/journal.pone.0028503.
Wu, C., and Morris, J.R. (2001). Genes, genetics, and epigenetics: a correspondence. Science *293*, 1103-1105. 10.1126/science.293.5532.1103.
Wu, D., Sunkel, B., Chen, Z., Liu, X., Ye, Z., Li, Q., Grenade, C., Ke, J., Zhang, C., Chen, H., et al. (2014). Three-tiered role of the pioneer factor GATA2 in promoting androgen-dependent gene expression in prostate cancer. Nucleic Acids Res *42*, 3607-3622. 10.1093/nar/gkt1382.
Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet *18*, 517-534. 10.1038/nrg.2017.33.
Xia, J., Sun, J., Jia, P., and Zhao, Z. (2011). Do cancer proteins really interact strongly in the human protein-protein interaction network? Comput Biol Chem *35*, 121-125. 10.1016/j.compbiolchem.2011.04.005.
Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res *37*, D105-110. 10.1093/nar/gkn851.
Xiao, L., Feng, Q., Zhang, Z., Wang, F., Lydon, J.P., Ittmann, M.M., Xin, L., Mitsiades, N., and He, B. (2016). The essential role of GATA transcription factors in adult murine prostate. Oncotarget *7*, 47891-47903. 10.18632/oncotarget.10294.
Xie, Y.H., Chen, Y.X., and Fang, J.Y. (2020). Comprehensive review of targeted therapy for colorectal cancer. Signal Transduct Target Ther *5*, 22. 10.1038/s41392-020-0116-z.
Xu, J., Wang, Y.Y., Dai, Y.J., Zhang, W., Zhang, W.N., Xiong, S.M., Gu, Z.H., Wang, K.K., Zeng, R., Chen, Z., and Chen, S.J. (2014). DNMT3A Arg882 mutation drives chronic myelomonocytic leukemia through disturbing gene expression/DNA methylation in hematopoietic cells. Proc Natl Acad Sci U S A *111*, 2620-2625. 10.1073/pnas.1400150111.
Yamashita, M., Ukai-Tadenuma, M., Kimura, M., Omori, M., Inami, M., Taniguchi, M., and Nakayama, T. (2002). Identification of a conserved GATA3 response element upstream proximal from the interleukin-13 gene locus. J Biol Chem *277*, 42399-42408. 10.1074/jbc.M205876200.
Yang, J., Lusk, R., and Li, W.H. (2003). Organismal complexity, protein complexity, and gene duplicability. Proc Natl Acad Sci U S A *100*, 15661-15665. 10.1073/pnas.2536672100.
Yang, Z., Jones, A., Widschwendter, M., and Teschendorff, A.E. (2015). An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. Genome Biol *16*, 140. 10.1186/s13059-015-0699-9.
Yokoyama, A., Kakiuchi, N., Yoshizato, T., Nannya, Y., Suzuki, H., Takeuchi, Y., Shiozawa, Y., Sato, Y., Aoki, K., Kim, S.K., et al. (2019). Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature *565*, 312-317. 10.1038/s41586-018-0811-x.
Yoon, H.G., Chan, D.W., Huang, Z.Q., Li, J., Fondell, J.D., Qin, J., and Wong, J. (2003). Purification and functional characterization of the human N-CoR complex: the roles of HDAC3, TBL1 and TBLR1. EMBO J *22*, 1336-1346. 10.1093/emboj/cdg120.
Yu, C., Luo, D., Yu, J., Zhang, M., Zheng, X., Xu, G., Wang, J., Wang, H., Xu, Y., Jiang, K., et al. (2021). Genome-wide CRISPR-cas9 knockout screening identifies GRB7 as a

driver for MEK inhibitor resistance in KRAS mutant colon cancer. Oncogene. 10.1038/s41388-021-02077-w.
Yu, K., Chen, B., Aran, D., Charalel, J., Yau, C., Wolf, D.M., van 't Veer, L.J., Butte, A.J., Goldstein, T., and Sirota, M. (2019). Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. Nat Commun *10*, 3574. 10.1038/s41467-019-11415-2.
Yu, Z., Jian, Z., Shen, S.H., Purisima, E., and Wang, E. (2007). Global analysis of microRNA target gene expression reveals that miRNA targets are lower expressed in mature mouse and Drosophila tissues than in the embryos. Nucleic Acids Res *35*, 152-164. 10.1093/nar/gkl1032.
Yuan, M., Huang, L.L., Chen, J.H., Wu, J., and Xu, Q. (2019). The emerging treatment landscape of targeted therapy in non-small-cell lung cancer. Signal Transduct Target Ther *4*, 61. 10.1038/s41392-019-0099-9.
Zamanighomi, M., Jain, S.S., Ito, T., Pal, D., Daley, T.P., and Sellers, W.R. (2019). GEMINI: a variational Bayesian approach to identify genetic interactions from combinatorial CRISPR screens. Genome Biol *20*, 137. 10.1186/s13059-019-1745-9.
Zentner, G.E., and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. Nat Struct Mol Biol *20*, 259-266. 10.1038/nsmb.2470.
Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L.D., and Ferretti, V. (2019). The International Cancer Genome Consortium Data Portal. Nat Biotechnol *37*, 367-369. 10.1038/s41587-019-0055-9.
Zhao, M., Kim, P., Mitra, R., Zhao, J., and Zhao, Z. (2016). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res *44*, D1023-1031. 10.1093/nar/gkv1268.
Zhao, Z., and Shilatifard, A. (2019). Epigenetic modifications of histones in cancer. Genome Biol *20*, 245. 10.1186/s13059-019-1870-5.
Zhu, M., Lu, T., Jia, Y., Luo, X., Gopal, P., Li, L., Odewole, M., Renteria, V., Singal, A.G., Jang, Y., et al. (2019). Somatic Mutations Increase Hepatic Clonal Fitness and Regeneration in Chronic Liver Disease. Cell *177*, 608-621.e612. 10.1016/j.cell.2019.03.026.