UNIVERSITY
OF SUSSEX

Exploiting patterns in genomic data for personalised cancer treatment and new target discovery

A thesis submitted to the University of Sussex for the degree of

Doctor of Philosophy

By: Graeme Benstead-Hume

September 2019

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Graeme Benstead-Hume

5 September 2019

Preface

The research presented in this thesis has been submitted for publication as follows:

Chapter 2

Benstead-Hume G, Wooller SK, Pearl FMG (2017) **Computational Approaches to Identify Genetic Interactions for Cancer Therapeutics**. Journal of integrative bioinformatics 14(3)

http://dx.doi.org/10.1515/jib-2017-0027

Author contributions: G.B.-H., S.K.W. and F.M.G.P. researched and wrote the paper.

Chapter 3

Benstead-Hume G, Chen X, Hopkins SR, Lane KA, Downs JA, Pearl FMG (2019) **Predicting synthetic lethal interactions using conserved patterns in protein interaction networks**. PloS Comput Biol 15(4): e1006888. https://doi.org/10.1371/journal.pcbi.1006888

Author contributions: G.B.-H., J.A.D. and F.M.G.P. designed the research. G.B.-H. And F.M.G.P. developed the algorithms and designed the website. G B-H coded the algorithms and the website resource. G.B.-H., J.A.D and F.M.G.P. conducted data analysis and interpretation and contributed to manuscript preparation. G.B.-H., X.C., K.A.L. and S.H. performed the experimental validation work.

Chapter 4

Benstead-Hume G, Pearl, FMG **Slorth: Validated and predicted synthetic lethal gene pairs with associated drug, disease and orthology data**. Submitted to

Database.

Author contributions: G.B.-H. Designed and developed software, processed and migrated data. G.B.-H. And F.M.G.P. wrote the paper.

Chapter 5

Benstead-Hume G, Wooller SK, Dias S, Woodbine L, Carr AM, Pearl, FMG **Biological network topology features predict gene dependencies in cancer cell lines. Submitted to systems biology.** Submitted to Molecular Systems Biology. Preprint available: https://www.biorxiv.org/content/10.1101/751776v1

Author contributions: G.B.-H., S.K.W, & F.M.G.P. conceived the project. G.B.-H. & S.K.W, designed the analysis, implemented the informatics & undertook the data analysis. S.D, L.W and T.C funded and performed experimental validation. G.B.-H., S.K.W, and F.M.G.P wrote the paper.

Chapter 6

Benstead-Hume G, SK Wooller, Downs JA, Pearl FMG **Defining signatures of arm-wise ploidy and their associated drivers in kidney cancers**. Int. J. Mol. Sci. 2019, 20(22), 5762; https://doi.org/10.3390/ijms20225762

Author contributions: F.M.G.P., J.A.D and G.B-H conceived the project and designed the analysis. G.B-H and S.K.W implemented the informatics and undertook the data analysis. G.B.-H. and F.M.G.P. wrote the paper.

Acknowledgements

A massive thank you to Dr. Frances Pearl for giving me the opportunity and the support to make all this happen. I have learned a huge amount from you and I consider myself very lucky that you gave me a chance! Thanks also to Prof. Jessica Downs for your enthusiastic encouragement and expert advice and the Medical Research Council for funding everything.

Sarah Wooller, thank you for being my best PhD friend and making this whole process fun (mostly). Also thank you for all the lunches, deeply insightful discussions and putting up with all of my maths questions. Soon we can be postdoc friends!

Thank you to all the lovely people that have brightened up the lab over the years especially my PhD buddies Dr. Yusuf Ali, Dr. Xiangrong (Tina) Chen, Dr. Hanadi Baeissa & Fahmida Banani. Thanks also to Dean Sumner, Harry Bowles and Rebbekka Narash Haley for your masterful masters work.

Thank you to Dr. Justin Reese for setting me on this path which isn't half as bad as you made out.

Love and thanks to Joyce and Alan Benstead who have supported me emotionally and financially even when the pay-off looked less than certain.

Thank you to all grandparents, including Pauline Hume, for the invaluable childcare and thank you to Peter Hume for all the proof reading and an exceptional eye for detail.

And finally all my love to Victoria Benstead-Hume, you're the best, and Harrison, thanks for all the nights that you didn't wake me up…

Abstract

In response to a global requirement for improved cancer treatments a number of promising novel targeted cancer therapies are being developed that exploit vulnerabilities in cancer cells that are not present in healthy cells. In this thesis I explore different ways of identifying the vulnerabilities of cancer cells, with the ultimate aim of providing personalised therapies to cancer patients on an individual basis.

I first investigate approaches that utilise the concept of synthetic lethality. Therapies that exploit synthetic lethality are suitable where a specific tumour suppressor has been inactivated by a cancer and an identified synthetic lethal (SSL) pair for that gene may be therapeutically targeted.

Mainly due to the constraints of the experimental procedures, relatively few human SSL interactions have been identified. Here I describe computational systems approaches for predicting human SSL interactions by identifying and exploiting conserved patterns in protein-protein interaction (PPI) network topology both within and across model species. I report that my classifiers out-perform previous attempts to classify human SSL interactions. Experimental validation of my predictions suggest they may provide useful guidance for future SSL screenings and ultimately aid targeted cancer therapy development.

All predictions from this study have been made available via a new online database that I designed, built and published.

As an extension to this approach I used similar network features to predict gene dependencies, otherwise known as acquired essential genes, in specific cancer cell

lines. Genetic alterations found in each individual cell line were modelled using the novel approach of removing protein nodes to reflect loss of function mutations and changing the weights of edges in each protein-protein interaction network to reflect gain of function mutations and gene expression changes.

I report that base PPI networks can be used to successfully classify human cell line specific gene dependencies within individual cell lines, between cell lines and even across tissue types. Furthermore, my personalised PPI network models further improve prediction power and show improved sensitivity to rarer gene dependencies, an improvement which offers opportunities for personalised therapy. In a therapeutic context these essential genes would be suitable as individual drug targets for each specific patient.

Finally, I analyse copy number variance and ploidy in a set of cancers from kidney patients. Using clustering algorithms I investigate patterns in cancer cell line arm-wise ploidy and identify factors that may be driving this genomic instability.

Abbreviations

AML	Acute Myeloid Leukaemia
AUC	Area Under the Curve
AUPR	Area Under Precision Recall Curve
BRCA	Breast cancer
Chr	Chromosome
CNV	Copy Number Variation
COAD	Colon Adenocarcinoma
COSMIC	Catalogue of Somatic Mutations in Cancer
DDR	Damage Response
DLBC	Diffuse large B-cell lymphoma
DNA	DeoxyriboNucleic Acid
dSLAM	Diploid based synthetic analysis on microarrays
E-MAP	Epistatic mini-array profiles
EBI	European Bioinformatics Institute
EGFR	Epidermal growth factor receptor
ENSP	Ensembl Protein
ENST	Ensembl Transcript
FDA	Food and drug administration
GBM	Glioblastoma
GDC	Genomic Data Commons
GLM	Generalised linear model
GO	Gene Ontology
GOF	Gain Of Function
HGMD	Human Gene Mutation Database
HNSC	Head-Neck Squamous Cell Carcinoma
ICGC	International Cancer Genome Consortium

KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney renal papillary cell carcinoma
KL	Kullback–Leibler
LAML	Acute Myeloid Leukemia
LGG	Low Grade Glioma
LOF	Loss Of Function
LUSC	Lung Squamous Cell Carcinoma
MESO	Mesothelioma
mRNA	Messenger RNA
NCBI	National Center for Biotechnology nformation
NCI	National Cancer Institute
NMF	Non-negative matrix factorisation
OG	Oncogene
OOB	Out of bag
OV	Ovarian cancer
PDB	Protein Data Bank
PPI	Protein-protein interaction
PRAD	Prostate Adenocarcinoma
RF	Random Forest
RME	Recurrent Mutually Exclusive aberrations
RNA	RiboNucleic Acid
RNAi	RNA interference
ROC	Receiver Operating Characteristic
SARC	Sarcoma
SDL	Synthetic dosage lethal
SGA	Synthetic genetic array
SIFT	Sorting Intolerant From Tolerant

SNP	Single Nucleotide Polymorphism
SSL	Synthetic lethal
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TI	Therapuetic index
TS	Tumour Suppressor
UCEC	Uterine Corpus Endometrial Carcinoma
UV	Ultraviolet radiation
WGS	Whole Genome Sequencing
YKO	Double yeast knockout mutant sets

Table of Contents

1 - Introduction

1.1 Overview

Cancer represents a major and rising global health burden, with over 12 million newly diagnosed cases per annum, and is responsible for more than 15% of the world's annual deaths. The development of new improved cancer therapies is frequently cited as an urgent unmet medical need (Varmus and Kumar, 2013).

Traditional treatments such as chemotherapy often exhibit a low therapeutic index (TI) due to the challenges presented by selectively targeting cancer cells whilst sparing normal cells (Muller and Milton, 2012). As such off-target damage of healthy cells is a common side effect of these therapies (Coates et al., 1983). Additionally, due to the homozygosity of cancer colonies and the cancer cell's accelerated mutation rate, drugs that appear effective at the outset of therapy can fail if even a small number of genes, and their resulting daughters, harbour a resistance to that compound (Holohan et al., 2013).

In response to these challenges cancer drug discovery now increasingly focuses on identifying and developing targeted therapies that promise both improved efficacy and therapeutic selectivity (Yap and Workman, 2012).These therapies effectively target specific genes, or their protein products, allowing for much higher specificity when targeting cancers with known genetic vulnerabilities (Schrank et al., 2018) Despite progress in this field targetted therapies are still unavailable many cancer patients and challenges such as resistance remain (Esplin et al., 2014)

1.2 Cancer

Cancer is a disease driven by the deregulated development and proliferation of otherwise normal cells. In the event that genetic alterations in a cell's DNA disrupt the processes that usually carefully regulate cell growth and maintenance they can result in the uncontrolled proliferation of the abnormal cell-line and ultimately lead to the metastasis of a cancer (Douglas Hanahan, 2000; Hanahan and Weinberg, 2011).

The development of cancer is generally progressive with affected tissues commonly transitioning through a number of steps, driven by genetic alterations, before they reach malignancy. The first steps in cancer progression are generally hyperplasia, where a tissues contain an excessive number of cells, or metaplasia, where a tissue's normal cells are partially replaced by cells that usually belong in other tissue types (Giroux and Rustgi, 2017). Tissues that that exhibit these properties are commonly termed benign. As these abnormalities develop cells may start to display dysplasia, changes in shape and size and the loss of differentiation as they lose the features and functionality usually found in cells that constitute their parent tissue (Zaini et al., 2018). Cells with dysplasia are often termed pre-malignant. The final steps toward full malignancy are neoplasm where the abnormal cells are able to create new growth outside of their predesignated tissue and invasion where these neoplasms enter blood or lymph vessels and colonise distant sites (Douglas Hanahan, 2000; Klein, 2008).

1.2.1 Alterations in driver genes are key to cancer progression

As cancer cells lose differentiation they generally exhibit a higher rate of genetic alterations than healthy cells due to the deactivation of various damage response and repair mechanisms (Pearl et al., 2015). The majority of these alterations, even those that effect coding regions of DNA, do not confer the cell with a selective advantage for growth. These alterations are generally known as passenger mutations (Pon and Marra, 2015). Conversely, genetic alterations that can be directly associated with tumorigenesis and cancer progression are commonly described as driver mutations. The genes that these driver mutations occur in are termed driver genes (Bailey et al., 2018). These driver genes can in turn be categorised as either oncogenes or tumour suppressors.

1.2.1.1 Oncogenes

Oncogenes, genes that are often involved in cell regulation, division and growth, can lead to tumorigenesis and cancer growth through gain of function mutations (Croce, 2008), genetic fusions, such as the BRC-ABL fusion (Advani and Pendergast, 2002; Weisberg et al., 2007), deregulation of gene expression (Sharma et al., 2009) or copy number variations (Sismani et al., 2015). These alterations may lead to uncontrolled expression of a usually carefully controlled protein that up-regulates the cell growth cycle (Anderson et al., 1992; Lynch, 1987). Oncogenes are generally said to be dominant as only one allele needs to be mutated to result in a selective advantage for the cancer cell (Pon and Marra, 2015).

RAS (Pylayeva-Gupta et al., 2011) and MYC (Dang, 2012), both examples of oncogenes, are commonly over expressed and exhibit gain of function mutations in the pathways associated with the hallmarks of cancer as discussed below.

1.2.1.2 Tumour suppressors

Tumour suppressors are associated with the loss of function. Tumour suppressors are genes that usually play a role of controlling cell replication, DNA damage response or programmed cell death, apoptosis (Gamudi and Blundell, 2010). Generally when both copies of a tumour suppressor are damaged the loss of function can lead to unregulated cell growth and tumorigenesis. Tumour suppressor mutations are said to be recessive as both alleles of a tumour suppressor gene must be inactivated to fully disable the gene and confer selective advantage (Yarbro, 1992).

BRCA1 and BRCA2, examples of tumour suppressors, commonly exhibit loss of function mutations in cancers and again both affect pathways associated with the hallmarks of cancer (Hansen, 2006; Miki et al., 1994).

1.2.1.3 Genetic mutations are permanent changes to DNA

The genetic alterations that drive cancers can broadly be split into mutations, where a cell’s DNA is permanently changed via base substitutions, insertions or deletions (Greenman et al., 2007; Martincorena and Campbell, 2015), and structural variants such as copy number variants (CNVs); where the amount of DNA in a cell is changed either through loss or gain of genetic material (Beroukhim et al., 2010; Valsesia et al., 2013). Changes in gene or protein expression and epigenetic factors are also important factors in the development of a cancer (Rodríguez-Paredes and Esteller, 2011; Sharma et al.,

2009) .

Genetic mutations are events where the sequence of an organism's original genetic material, DNA, is changed. This can occur in the germ-line, where the mutation occurs in the gametes of an organism or as somatic mutations where the mutation occur in the cells of a developed individual (Martincorena and Campbell, 2015). Genetic mutations take many forms, from single nucleotide polymorphisms (SNPs) where a single base of DNA is changed to another base (Batra et al., 2014), insertion and deletions of bases, collectively known as indels (Sehn, 2014) and translocations where sequences of DNA are removed from their origin and situated elsewhere else in the genome (Bunting and Nussenzweig, 2013; Nambiar et al., 2008).

Silent mutations change the DNA in such a way that the resulting amino acids remain functionally unchanged. Silent mutations generally cause no damage. Missense mutations result in the original amino acid being substituted by another and may cause some change in function of the resulting protein (Adzhubei et al., 2010; Vaser et al., 2016). Nonsense mutations result in a stop codon and ultimately a truncated protein product. These usually result in a complete loss of protein product as a result of nonsense mediated decay (Asiful Islam et al., 2017; Wen and Brogna, 2008).

Insertions and deletions (indels) add or remove a small number of bases to or from the original DNA sequence (Sehn, 2014). While some indels, those with a length divisible by three, lead to the gain or loss of amino acids in the resulting protein product more often indels lead to a frameshift. A frameshift is where the reading frame, read in groups of three bases, is shifted by either one or two bases leading to the alteration of all codons subsequent to the mutation. Frameshifts often result in nonsense mediated decay which leads to degradation of the transcribed mRNA of the altered gene (Streisinger et al., 1966).

1.2.1.4 Mutations are caused by a wide range of factors

Mutations are generally caused either by damage from exogenous factors, known as carcinogens (Ames et al., 1973), or endogenous factors such as tissue inflammation (Ames and Gold, 1991).

There are a range of external factors that may lead to mutations in a cell's DNA. Common examples of carcinogens include various carcinogenic chemicals from food (Goldman and Shields, 2003) or environmental particles in smoke (DeMarini, 2004) or dust (Huang et al., 2011; Rekhadevi et al., 2009) as well as exposure to certain types of radiation (Behjati et al., 2016), by-products of tissue inflammation and hormone imbalances. Certain viruses have also been implicated in tumorigenesis (Mesri et al., 2014).

Tobacco smoke contains a number of chemicals including Arsenic, Benzene and Formaldehyde all of which have been proven to be carcinogenic and commonly lead to cancers such as lung adenocarcinoma (Lehman et al., 1991; Yarbro, 1992). Many foods also contain chemicals known to be carcinogenic such as heterocyclic amines and polycyclic aromatic hydrocarbons found in cooked meats and acrylamide in burnt toast (Goldman and Shields, 2003).

Ultraviolet radiation (UV), present in sunlight, is known to cause DNA damage which, when paired with mistakes in DNA repair mechanisms can lead to disrupted cellular processes and potential tumorigenesis (Fitzpatrick and Sober, 1985). Ionizing radiation such as X- and Gamma-radiation which arise from radioactive decay and which are commonly used in medical imaging are also mutagenic (Behjati et al., 2016).

More generally a range of lifestyle factors are associated with an increase an individual's chances of developing cancers such as diets that are high in fat or salt and those low in fibre (Kushi and Giovannucci, 2002; Tsugane, 2005). Alcohol consumption, especially in conjunction with tobacco, has also been associated with increased cancer risk (Blot et al., 1988). For a review of the association between lifestyle and cancer see Nagahashi et al. (Nagahashi et al., 2018). Finally age is a significant factor in the onset of cancer due to a life-time's accumulation of mutations.

Many of the above factors promote inflammation and either directly cause mutations in the host's DNA or otherwise disrupt cell function leading to further mistakes in cellular maintenance such as improper DNA repair or duplication.

1.2.1.5 Copy number variance is the loss or gain of genetic material

Copy number variation, a form of structural variation, is where sections of an individual's DNA occur more or less often than would be expected in healthy cells. In humans, where DNA is naturally diploid, a copy number variation may include sections of DNA that only occur once or those that occur three or more times (Hastings et al., 2009). Copy number variations commonly occur as a result of mistakes in mitosis or non-homologous end joining which lead to genome instability (Sismani et al., 2015).

CNVs may play a role in tumorigenesis by either disrupting or altering dosage of oncogenes or tumour suppressors that are contained in the CNV (Shlien and Malkin, 2009). Approximately 40% of known cancer genes are disrupted by CNVs (Almal and

Padh, 2012) and it is estimated that of all known CNVs 40% contain genes implicated in cell growth and metabolism (Conrad et al., 2010).

Aneuploidy, where entire copies of chromosomes are gained or lost, can lead to multiple copies of an oncogene in the case of a gain of chromosomes or a loss of heterozygousity of a tumour suppressor in the case of a lost chromosome (Gordon et al., 2012; Orr et al., 2015; Rajagopalan and Lengauer, 2004).

1.2.1.6 Gene expression is commonly deregulated in cancers

Tightly controlled gene expression is an important aspect in the regulation of a healthy cell. Long term gene silencing is provided via epigenetic changes, these epigenetic modifications are mitotically heritable and so form a stable part of genetic information that distinguishes fully differentiated cells. Other regulation varies depending on the circumstances of the cell such as cell cycle phases and exogenic factors (Riley and Anderson, 2011; Sharma et al., 2009).

Commonly the expression of a gene is physically regulated via histone modifications and methylation of CpG nucleotides in promoter regions and other control regions that enhance or insulate transcription (Conrad et al., 2010). These modifications can lead to the inhibition of transcription though a range of mechanisms, for example they may block DNA polymerase from binding to promoter sites and they may promote the formation of inactive chromatin sites. There are also likely other mechanisms that are not well understood (Weiderpass, 2010).

The deregulation of gene expression and methylation plays a direct role in tumorigenesis and cancer progression. Three mechanisms of gene expression deregulation commonly found in cancer cells. These include the direct silencing of tumour-suppressors via hyper-methylation, for example *TGFBI*, *SPARC*, *RBP1*, *GPX3* are often found silenced in multiple myeloma (Jones, 2012) and the widespread hypomethylation of DNA which leads to genomic instability (Gokul and Khosla, 2013). Finally the wide-spread deregulation of gene expression caused by changes in the methylation of regulatory regions within miRNA coding genes leads to changes in the transcription of miRNA and consequently the transcription of genes normally regulated by the miRNA. This may lead to the knock-out of tumour suppressors or increased expression of oncogenes (Kaiser et al., 2013).

1.2.2 The hallmarks of cancer

In their seminal papers ‘The Hallmarks of cancer’ and ‘The Hallmarks of cancer: the next generation’ (Douglas Hanahan, 2000; Hanahan & Weinberg, 2011) Weinberg and Hanahan describe the changes required in a cell for it to develop into a fully invasive cancer cell (Figure 1.1). As a tumour incrementally acquires these traits through the genetic alterations discussed above it moves closer to developing into a fully invasive cancer.

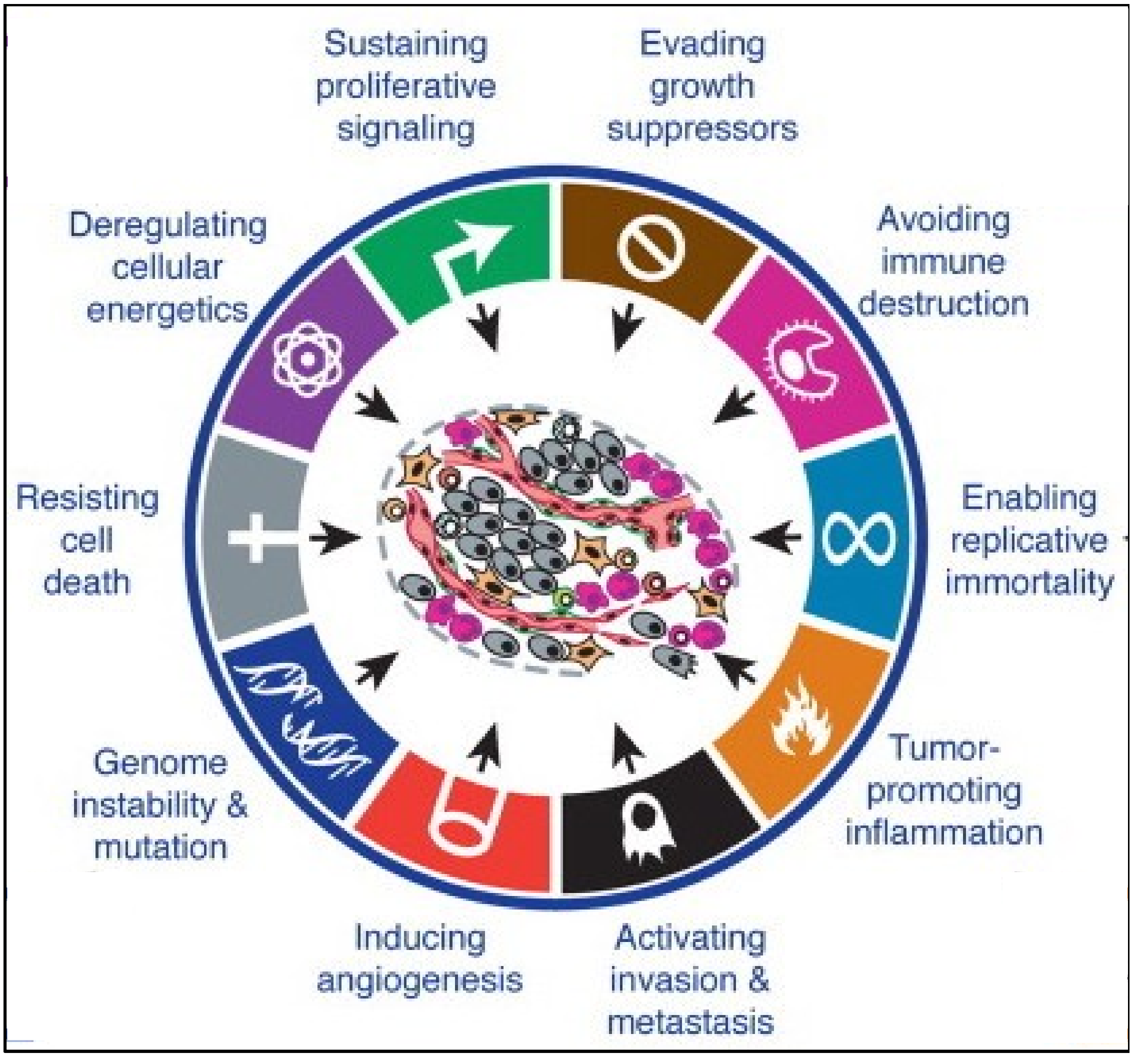

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

Figure 1.1: The 10 hallmarks of cancer. Weinberg and Hanahan describe the changes required in a cell for it to develop into fully invasive cancer; sustaining proliferative signalling, evading growth suppressors, avoiding immune destruction, enabling replicative immortality, tumour-promoting inflammation, activating invasion and metastasis, inducing angiogenesis, genome instability and mutation, resisting cell death and deregulating cellular energetics. Figure sourced via (Hanahan and Weinberg, 2011).

1.2.2.1 Resisting cell death (Evading Apoptosis)

Mutational events are common in somatic cells and so, in order to ensure healthy tissue and avoid potential tumorigenesis, there exist a number of molecular processes that check for damage, cell stress or changes in function and which force the damaged cell to enter programmed cell death, also known as apoptosis, when a cell is beyond repair (Elmore, 2007).

DNA damage is a prerequisite to tumorigenesis and all of the traits of cancers listed here. The disruption of either the pathways that sense DNA damage or those that act on the signals to further drive apoptos is a key step in the progression of cancer (Hengartner, 2000).

A range of genes are complicit in the pathways that drive damaged cells towards apoptosis. A well-studied protein involved in apoptotic pathways; p53 is the protein product of TP53, a gene commonly knocked-out in cancer tissues (Fridman and Lowe, 2003).

1.2.2.2 Sustaining proliferative signalling (Self-sufficiency in growth signals)

The propagation of healthy cells is a tightly controlled process driven via intracellular and extracellular signals.

In order for cancer to progress it must bypass the requirement for extra-cellular growth signals. Tumours achieve this by subverting a number of different pathways to either create their own growth signals, heterotypic signalling, or to greatly reduce the need for exogenous growth signals (Slamon et al., 1987; Witsch et al., 2010; Yarden and Ullrich, 1988).

1.2.2.3 Insensitivity to anti-growth signals (Evading growth suppressors)

Normal cells are subject to a number of anti-proliferative signals used to restrict proliferation in the case of unsuitable extracellular conditions such as overcrowding or the detection of errors in the cell cycle. This is generally achieved by forcing cells into G0 phase and into quiescence until the issue is resolved. In cases where differentiation is detected cells are ultimately forced into a post-mitotic state where a cell is no longer able to proliferate. (Datto et al., 1997; Hannon and Beach, 1994).

1.2.2.4 Sustained angiogenesis (Inducing angiogenesis)

As a prerequisite to continued growth a tumour must increase the amount of nutrients and oxygen delivered, as well as the means of removing waste products, via blood vessels.

In normal tissues the arrangement of blood vessels is carefully regulated and immutable by neighbouring cells due to angiogenesis inhibitors. Cancer cells must develop sustained angiogenic ability by removing angiogenesis inhibitors and through the up-regulation of angiogenesis promoting proteins (Bouck et al., 1996; Hanahan and Folkman, 1996).

1.2.2.5 Limitless replicative potential (Enabling replicative immortality)

As well as the active signalling that promotes or represses proliferation most normal mammalian cells also exhibit hard limits on replicative potential, typically around 60-70 duplications. This limit is a result of the progressive, the natural shortening of telomeres the sections of DNA at the end of chromosomes that provide protection from end-to-end chromosomal fusions. The majority of cancer cell lines mitigate this limit by up-regulating the expression of the telomerase enzyme, the proteins that maintain telomere length (Bryan and Cech, 1999; Shay and Bacchetti, 1997).

1.2.2.6 Tissue invasion and metastasis (Activating invasion and metastasis)

For a tumour to be defined as an invasive cancer it must first start to grow daughter cells that are able to migrate out of the original tissue. These cells will then traverse blood or lymph systems and eventually create separate colonies in new areas of the patient's body.

To metastasise a tumorous cell must first break the bonds of the extracellular matrix and have a means of traversing barriers such as epithelial walls (Jiang et al., 2015).

1.2.2.7 Genome instability and mutation

Genetic aberrations are very common in all cells due to issues such as mistakes in cell duplication, genomic instability and exogenous factors. Fortunately healthy cells have a suite of DNA damage detection and repair mechanisms which are usually able to identify and fix issues before they can cause serious problems.

As discussed previously cancer is a genetic disease in that it is driven by aberrations and cancer cells that tolerate mutations and genetic instability have a selective advantage over cells that do not. As such mutations and instabilities are common in most cancers and the rate of these aberrations is further accelerated in tumorous cells through disruptions in the cell's DNA damage response and repair functionality as discussed above (Jackson and Bartek, 2009; Lord and Ashworth, 2012; Negrini et al., 2010; Tlsty et al., 1995).

1.2.2.8 Reprogramming energy metabolism (Deregulating cellular energetics)

To maintain an accelerated rate of proliferation many cancer cells switch from aerobic respiration to the less efficient but significantly faster anaerobic respiration to produce ATP to fuel metabolic reactions.

While cancer cells using anaerobic respiration effectively require 20 times as much glucose to run these reactions the result is that these cells can produce ATP at a rate almost a hundred times faster than a normal cell. Additionally, anaerobic respiration produces other by-products that further fuel accelerated proliferation (Lunt and Vander Heiden, 2011)

1.2.2.9 Avoiding immune destruction

The immune system is an important factor in an organism's defence against cancer. In most cases an abnormal cell will flag itself as such via antigens and will be targeted for controlled destruction via T cells, natural killer cells or macrophages. In many cancers proteins associated with antigens can be under-expressed leading to a disruption in antigen presentation and abnormal cells that are harder to identify for destruction (Rouas-Freiss et al., 2003).

1.2.2.10 Tumour promoting inflammation

As well as avoiding destruction by the immune system some tumorous cells use mechanisms of the immune response, most commonly inflammation, to further accelerate growth.

Inflammation is often exploited in cancers to accelerate proliferation, and indirectly promote angiogenesis to provide immunosuppressive support and to degrade the surrounding extra cellular matrix, ultimately aiding metastasis .

1.2.3 Cancer therapy

By necessity cancer therapies must attack the aberrant cells once a tumour is discovered. Traditional chemotherapy regimes commonly employ cytotoxic agents to cause damage or structural changes to DNA that fast replicating cells are unable to repair before completing a cell cycle. This unrepaired damage results in the cells inability to replicate DNA interfering mitosis indirectly leading to DNA damage response and the activation of apoptotic pathways (Hennequart et al., 2017; Woods and Turchi, 2013). These therapies are designed based on the logic that cancer cells divide faster than most healthy cells which affords some target specificity (DeNardo et al., 2010; Grivennikov et al., 2010; Karnoub and Weinberg, 2016; Qian and Pollard, 2010).

Unfortunately, these therapies also affect other “healthy” but rapidly dividing cells leading to significant damage in unintended targets. This off-target damage commonly results in the trademark side-effects of cancer therapy such as gastrointestinal upset and hairloss (Coates et al., 1983)

A therapeutic index (TI) is the ratio of the dose of a therapeutic agent that causes toxicity (that is a lethal dose in 50% of subjects, LD50) to the amount that causes the desired therapeutic effect (Effective dose in 50% of subjects, ED50). As such a drug with a larger TI is favourable to drug with a low TI.

Standard chemotherapies often have a low TI due to the challenge presented by selectively targeting cancer cells whilst sparing normal cells (Muller and Milton, 2012). Furthermore, due to cancer cells' predisposition to acquire mutations, a drug that seems effective at the outset of therapy may well be rendered ineffective if even a single cell, and its resulting daughters, gain resistance to that compound (Holohan et al., 2013). In response to these challenges a number of targeted therapies designed to increase TI are in development or have in some cases been approved (Santos et al., 2016).

As of 2016, 85 of the available 154 cancer drugs licensed by the FDA, were targeted therapies designed to target the genes that directly drive cancer (Santos et al., 2016). Many targeted anti-cancer drugs work by directly inhibiting activated oncogenes, particularly proteins that are nuclear receptors or those that contain protein kinase domains (Iorio et al., 2016; Nguyen et al., 2017; Shawver et al., 2002). Dabrafenib, which has been approved for the treatment of late-stage melanoma, target the constitutively activated kinase oncogene BRAF V600E. Whilst gefitinib and erlotinib, licensed for the treatment of lung cancer, targets the EGFR tyrosine kinase (Lindeman et al., 2013; Shepherd et al., 2005; Stinchcombe and Socinski, 2008; Thatcher et al., 2005).

A substantively different approach is needed to provide therapies aimed at controlling the damage done by inactivated tumour suppressor genes. It is not usually feasible to repair the protein products of these genes particularly if they are inactivated by truncation, although there are on-going attempts to reactivate or restore function to a small subset of p53 missense mutant proteins (Burgess et al., 2016; Hoe et al., 2014).

To exploit genetic interactions therapeutically, the genetic defects in an affected pathway must be combined with a pharmacologically induced defect in a compensating pathway. Synthetic lethality (SSL), discussed in more detail below, is well suited for targeting deactivated tumour suppressors (Hartwell et al., 1997). SSL causes cell death as a result of one gene being genetically inactivated by mutation (loss of function (LOF), the tumour suppressor) and another being inactivated by a drug target. While synthetic dosage lethal interaction can be used for targeting cancer cells with over-expressed oncogenes (Megchelenbrink et al., 2015). SDL causes cell death as a result of one gene being genetically activated (gain of function (GOF), the oncogene) and another being inactivated (LOF, the drug target).

Targeted therapies that exploit these genetic interactions may provide a significantly improved therapeutic index compared to standard chemotherapies (McLornan et al., 2014).

1.2.3.1 Resistance to cancer drugs

Natural selection drives cancer in that cells that are able to out-compete their neighbours are effectively selected for continued growth and replication. While cancer therapies are often very effective against the majority of target cancer cells occasionally a small number of those cells will have acquired a mutation that renders them tolerant or fully resistant to the intervention. In these cases the surviving cells many continue to grow into a fully resistant colony without the competition from the cells affected by the therapy (Holohan et al., 2013).

Cancer cells can achieve drug resistance in a number of ways including releasing proteins that modify or degrade the drug in vivo, altering the drug’s molecular target, reprogramming the DDR to repair drug damage and inhibiting cell death.

Drug resistance is one of the primary challenges in cancer medicine. The exploitation of genetic interactions and synthetic lethality may offer opportunities to develop therapies less prone to resistance (Porcelli et al., 2012).

1.3 Genetic interactions and synthetic lethality

Genetic interactions and, specifically synthetic lethal interactions, feature heavily in the proceeding chapters due to the therapeutic opportunities that they present. Here we provide an overview while Chapter 2 provides a full review of genetic interactions, their role in cancer therapeutics and attempts to identify them in different organisms.

Genetic interactions are a combination of genes where a change in the regulation of both genes concurrently results in a more extreme phenotype than what we would expect from the independent change of two unrelated genes (Costanzo et al., 2010). Genetic interactions are often categorised as either negative or positive genetic. Negative genetic interactions are where a combination of genetic alterations, such as mutations, result in a less viable phenotype than expected based on the sum of the two individual alterations. Positive genetic interactions are where the resulting phenotype of two concurrent gene alterations is less severe than expected.

Synthetic lethal interactions, a class of negative genetic interaction, are where two concurrent deleterious gene alterations cause cell death or a notably detrimental phenotype while individual mutations in either gene alone leave the cell viable. Synthetic dosage interactions (SDL), another class of negative genetic interaction, is where the up-regulation of one cell combined with a deleterious alteration of the other results in cell death or reduced viability (Figure 1.2) (Nijman, 2011).

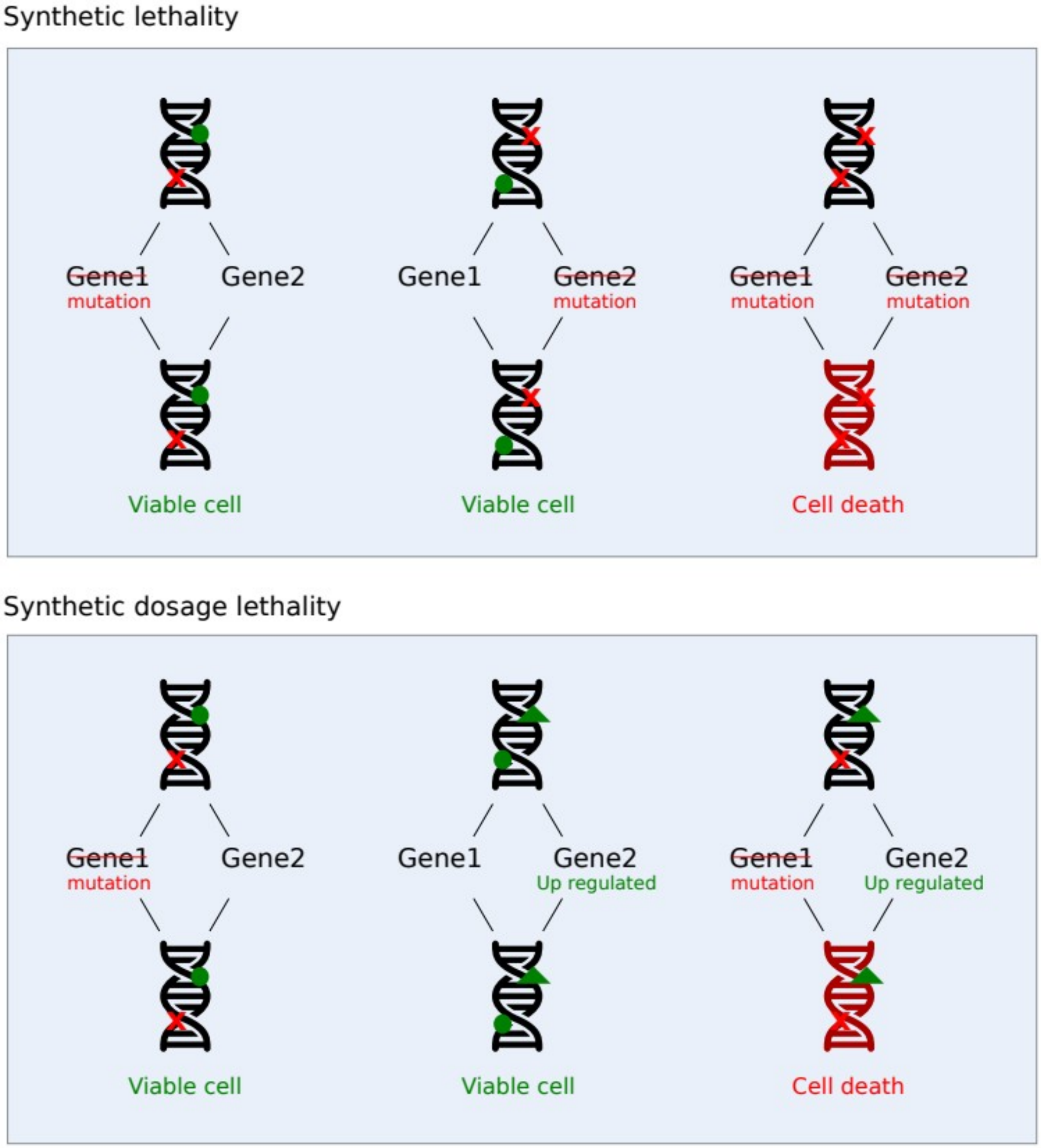


Figure 1.2 A schematic of negative genetic interaction types. In the example depicting SSL a cell remains viable when either gene in the interacting pair is deleteriously altered individually, however when both genes are altered the cell loses viability. In the case of SDL the cell loses viability when one gene is deleteriously altered at the same time that the other is up regulated.

1.3.1 Synthetic lethality in the clinic

Negative genetic interactions and more specifically SSL interactions present a novel opportunity to target genes that are not currently directly druggable. Tumour suppressors, for example, rarely make suitable therapeutic targets as the majority of drugs are not suited to restoring gene functionality.

As an alternative to directly targetting these genes, synthetic lethal interactions may be exploited in cases where a tumour suppressor is mutated in a tumour and that mutated tumour suppressor shares a synthetic lethal interaction with a gene that may be therapeutically targeted.

Therapies that exploit synthetic lethal interactions are already in use in the clinic. One of the more successful of these being the PARP inhibitor Olaparib which is used for certain breast and ovarian cancer patients. This PARP inhibitor exploits the synthetic lethal interaction between the BRCA genes and PARP1. The BRCA genes are both fundamental in DNA damage repair pathways, most notably homologous recombination, and both commonly mutated in cancers. PARP1, is also implicated in DNA repair pathways involving single-strand breaks and base excision repair. When both a BRCA gene and PARP1 are concurrently suppressed unrepaired DNA damage accumulates leading to genetic instability and cell death (Figure 1.3) (Bryant et al., 2005; Lord and Ashworth, 2017).

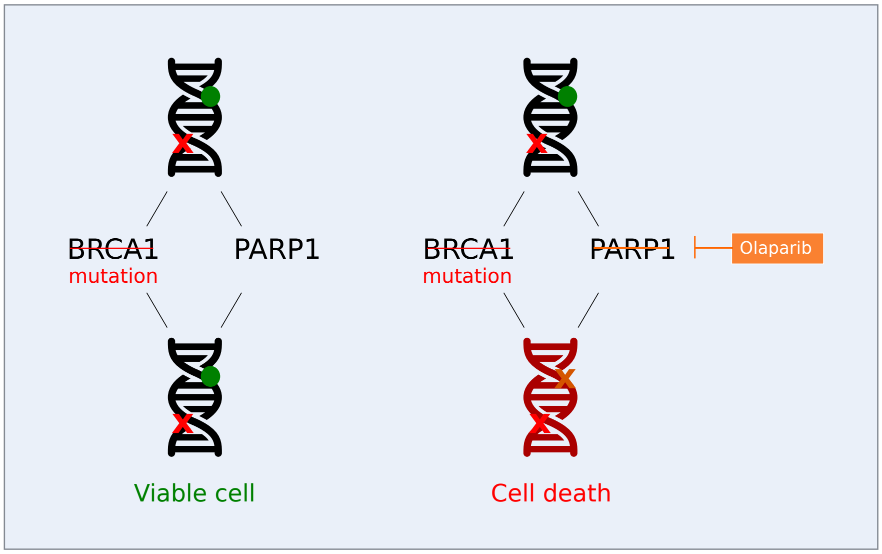


Figure 1.3 A schematic of how Olaparib, a PARP1 inhibitor, exploits the SSL interaction between BRCA1 and PARP1. In this case BRCA1 is mutated by a cancer. BRCA1 shares an SSL interaction with PARP1 which is therapeutically targetted leading to SSL and cell death.

1.3.2 Experimental validation of SSL

Synthetic Lethality was first described in 1922 by Calvin Bridges following a study crossing Drosophila melanogaster. Contemporary SSL studies have classically focused on crossing eukaryotic model organisms with increasing sophisticated techniques that allow researchers to develop hybrid genomes and to screen them using gene silencing techniques such as RNA interference (RNAi). More recently high throughput approaches to finding genetic interactions in model organisms have been developed based broadly around three distinct platforms; synthetic genetic array (SGA) (Tong et al., 2001b), diploid based synthetic analysis on microarrays (dSLAM) (Pan et al., 2007) and epistatic mini-array profiles (E-MAP) (Collins et al., 2010). These methods are further discussed in Chapter 2.

Although it was at one time hypothesised that SSL pairs would be conserved across species in orthologous genes (Wu et al., 2014), it has since been found that often SSL interactions are not conserved between lower eukaryotes and their human orthologous equivalents (Boucher and Jenna, 2013). As a result, although SSL data for model organisms can teach us a lot about gene function and pathway interaction, we cannot rely on previous work on model organisms to inform us of SSL relationships in human cell lines.

Classically SSL discovery in humans has been a ‘hypothesis driven’ process of predicting SSL interactions based on proven associations, often related to loss of particular cell cycle checkpoints or pathways related to those of known tumour suppressors, and subsequent clinical trials. However, with the increasing availability of genetically modified human cell lines and high throughput genetic screening methods that combine RNAi screens with libraries of small molecule inhibitors, an increasing number of human SSLs are being identified (Barbie et al., 2009; Berns et al., 2004; Luo et al., 2009a; Scholl et al., 2009; Turner et al., 2008).

Despite the potential therapeutic opportunities the exhaustive experimental identification of human SSL interactions is not currently tenable due primarily to the sheer number of gene pair screens required (You et al., 2010).

Instead we turn to computational approaches in order to predict potential SSL pairs that may be used to guide future screening.

1.3.3 Predicting synthetic lethal interactions

1.3.3.1 Computational identification of genetic interactions

A systematic approach to inferring genetic interactions has become increasingly popular in the past decade. The ever-growing amount of screening data available has paved the way for more sophisticated computational techniques employing statistical and machine learning. These in silico models have proved significantly cheaper and faster to implement compared to traditional screening methods and have demonstrated impressive levels of accuracy when predicting genetic interactions (Jacunski et al., 2015; Madhukar et al., 2015a; Paladugu et al., 2008; Wong et al., 2004a; Zhong and Sternberg, 2006).

The most prevalent models used to predict genetic interactions use biological network data, gene ontology, expression level data, and orthology or evolutionary data. Historically these studies have generally focused on model organisms, due to the availability of the data for these more easily studied organisms. However, more recently systems biology data for humans is becoming more widely available and more complete (Lehne and Schlitt, 2009).

For a full review of SSL and contemporary computational studies see Chapter 2.

1.3.3.2 Synthetic lethal databases

Considering the excellent opportunities that genetic interactions, and more specifically synthetic lethality, present in cancer drug discovery, relatively few publicly available databases exist to share experimentally validated or computationally predicted human synthetic lethal pairs.

Two prominent databases exist for yeast genetic interaction data including cellmap (Dallago et al., 2018; Usaj et al., 2017) and the Saccharomyces Genome Database (Cherry et al., 1998).

For human genetic interactions data BioGRID (Stark, 2006) is a well curated source for experimentally validated genetic interactions including SSL and SDL interactions for a range of organisms. The majority of the SSL data used in the following studies use SSL data generated through synthetic genetic array (SGA) screens (Tong et al., 2001a) via BioGRID.

SynlethDB (Guo et al., 2015) previously published a range of validated and predicted genetic interactions for a number of organisms including humans but the database not currently maintained.

1.4 Machine learning and classification

Machine learning encompass a range of techniques that enable computer programs to improve their performance at a given task given experience. Commonly machine learning is used to learn from observed data (experience) in order to improve prediction or classification (performance) of future observations (task) without explicit programming. This kind of machine learning has become an essential tool in the analysis of bio-medical data (Angermueller et al., 2016; Tarca et al., 2007; Zhang and Rajapakse, 2008).

The recent growth in the availability, quantity and complexity of genomic, proteomic and interaction data due to increasingly sophisticated sequencing and screening protocols has demanded equally sophisticated approaches for analysis. As datasets continue to grow larger than might be realistically managed and analysed using traditional statistical techniques, machine learning is proving a versatile alternative for extracting valuable insight from raw experimental observations (Zhang and Rajapakse, 2008).

Machine learning can broadly be divided into two sub-disciplines, supervised and unsupervised learning. Supervised learning can be further subdivided into either regression or classification models.

1.4.1 Supervised learning

During supervised learning a model will learn from training data that includes both features, X and observed outcomes or labels, Y (Kotsiantis, 2007).

While there are many different supervised learning algorithms the studies presented here have focused on decision trees and, more specifically, random forest classifiers, as they encode conditionality which suits biological data (Qi, 2012).

1.4.1.1 Decision trees and Random Forest classifiers

A decision tree is a model of decisions and their respective consequences in the form of a branching graph (Figure 1.4). Decision trees are generally easily interpretable and encode conditionality making them a suitable model for a range of biological processes.

A random forest classifier is an ensemble machine learning method that takes a consensus from a number of individual decision trees built using random subsets of the original feature space (Yamaoka, 2012).

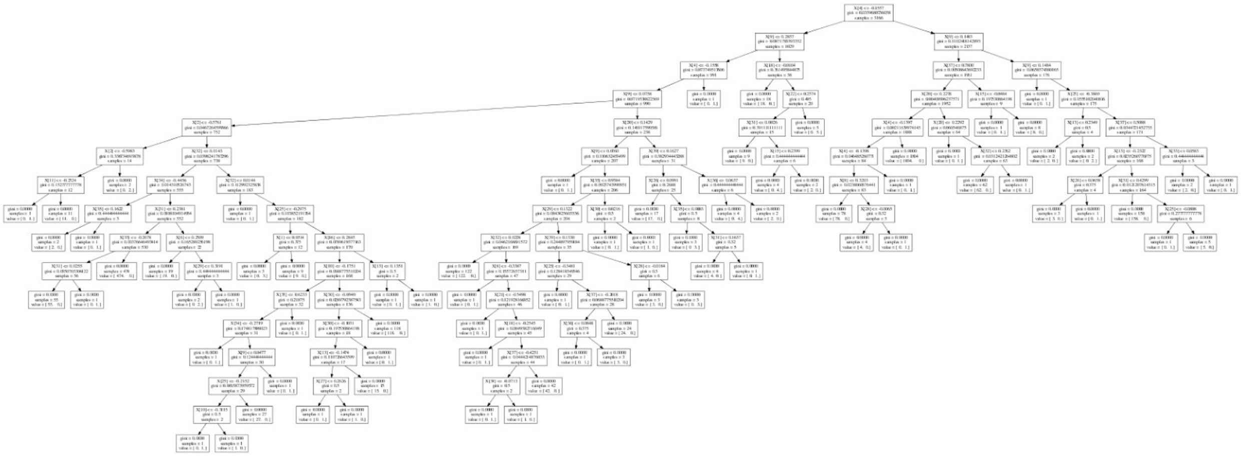


Figure 1.4 A decision tree with branching probabilities

Decision trees are grown using a simple algorithm:

1. Choose best attribute for root node A
2. For each possible value of A create a new child node
3. For each child node, stop if node is pure (if the observations defined by that node are all of one class) or, otherwise, recursively split into further child nodes

To create the most efficient decision tree we must choose the best attribute to split at each branch. To do this we must measure the purity of the labelled classes in the resultant child nodes for each potential split. One way to measure purity across multiple classes by taking a measurement of the uncertainty of a class in a subset of examples. This measurement is known as entropy.

Equation 1.1.

$$H(S) = -p(positive)\log 2\, p(positive) - p(negative)\log 2\, p(negative)$$

Where S is the subset of examples and positive and negative can either be classes in a binary feature or binarised values created by applying thresholds to continuous features.

In a random forest classifier a given number of trees are grown using a random sample of roughly two thirds of the available training data and with m features (m is commonly set to be the square root of the total number of available features) which are selected at random from the full feature set. The remaining one third of data is set aside to act as out of bag (OOB) data for validation. The holdout data is used to calculate the misclassification or OOB error rate and the overall OOB error rate for the classifier is calculated by taking the aggregate error across all trees (Oshiro et al., 2012; Segal, 2004).

At classification each tree in the random forest is queried based on a given observation's feature values and each tree's resulting output acts as a 'vote' for the predicted class. The random forest classifier selects the class that receives the most votes in this way. If for example class 1 receives votes from 150 trees and class 2 receives votes from 50 trees the random forest classifier will classify class 1 with a 0.75 probability (Strobl et al., 2007; Yamaoka, 2012).

1.4.2 Unsupervised learning

Unsupervised learning clusters observations into classes using feature data X only. As such an unsupervised model does not need labelled data Y to learn (Francis, 2014)

The unsupervised learning techniques presented in the chapters below have focused on non-negative matrix factorisation, made popular in biological contexts by Nik-Zainal et al. (Nik-Zainal et al., 2012; Stark, 2006)

1.4.2.1 Non-negative matrix factorisation

Non-negative matrix factorisation (NMF) is a multivariate analysis tool used in a range of fields for easily interpretable decomposition, dimensionality reduction to a given number of components and clustering (Lazar et al., 2009).

Essentially NMF decomposes a feature matrix V into two descriptive matrices, a basis, W, which describes the feature composition of each component and a coefficient, H which describes the component composition of each sample in the original matrix .

If we take the matrix, V, our initial data, to be the product of matrices W and H as such:

V = WH

We can optimise W and H by minimising the error function:

Equation 1.2

$$Err(W,H)=min(W,H \| V-WH \|)$$

Where V is the original matrix of data, W is the basis matrix and and H is the coefficient matrix as described above.

In the following studies we approximate W and H using the Kullback–Leibler (KL) divergence algorithm. To optimise W and H, KL divergence calculates how well WH approximates V by measuring the resultant cross entropy minus the entropy between the two matrices as an error:

Equation 1.3

$$\sum_{ij}(V_{ij}\log(\frac{V_{ij}}{WH_{ij}})-V_{ij}+WH_{ij})$$

with W and H updated over a number of iterations to minimise this error (Guo et al., 2015; Lee and Seung, 2001).

1.4.3 Machine learning best practices

1.4.3.1 Generalisation and validation

The primary challenge of machine learning is to train a model that not only successfully fits the given training data but also generalises well to new, previously unseen observations. A model that performs well on the training set but performs poorly on unseen data is said to be biased or over-fitted (Brownlee, 2016; D1Etterich, 1995)

To measure this generalisation we generally hold aside a subset of our labelled data to be used as a final test of our model (Martin, 2016).

Additionally, in order to tune our models, we must further divide our remaining training data into training and validation sets. This validation data can be used to measure predictive power across different models and associated hyper-parameters. Where dataset are relatively small this training / validation split can result in diminished sample sizes which can impact the generalisation of a model. In these cases we can use cross-validation. Cross validation is where a training set is split in to x segments, validation is performed x times with each segment taking a turn to act as the validation data while the other x-1 segments are used as training data. This approach enables the full use of the remaining training data (Bengio and Grandvalet, 2004; Park and Kim, 2012).

Ideally holdout test data should only be used once to evaluate model performance to avoid bias leaking back into the model via posthoc amendments (Martin, 2016).

1.4.3.2 Feature importance

Measuring the importance of features in a trained model can provide insight into underlying mechanisms that contribute to an observed class. The models in the studies presented here are used to calculate feature importance by measuring the mean decrease in accuracy at validation with each feature systematically withheld across all tree permutations in a random forest (Iguyon and Elisseeff, 2003; Saeys et al., 2007) (Menze et al., 2009; Strobl et al., 2007).

1.4.3.3 Feature scaling

Feature scaling is important in a number of classifiers that learn by ascribing weights to features such as linear regression, neural networks and support vector machines. Without scaling these models tend to be biased towards features with large values and as such features are often normalised so that feature values are on the same scale (Jacunski et al., 2015).

Generally feature scaling is not a major concern when using random forest classifiers as features are not weighted in the same way. However in the projects presented the random forest classifiers are used to predict on data across datasets, i.e. we train a model using human biological data and use that model to predict classes in yeast data. In this case feature scaling is particularly beneficial to predictive success (Jacunski et al., 2015; Nik-Zainal et al., 2012).

1.5 Protein – protein interaction data

Also known as physical interaction data or the interactome, protein - protein interaction (PPI) data is a map of how an organism's proteins are functionally associated. In the studies presented here we use experimentally validated PPI data sourced via STRING (von Mering et al., 2005). STRING itself features experimentally validated PPI pairs extracted from a range of primary databases including BIND (Bader and Hogue, 2003)DIP (Xenarios, 2000), GRID (Breitkreutz et al., 2002), HPRD (Peri, 2004), IntAct (Hermjakob, 2004), MINT (Peri, 2004), and PID (Schaefer et al., 2009).

1.5.1 Experimental identification of PPI data

The primary PPI databases listed above include evidence from many different experimental methods all with different levels of reliability. The most reliable of these methods included x-ray crystallography (Bunaciu et al., 2015; Huxford, 2013), Nuclear Magnetic Resonance Spectroscopy (NMR) (Kalbitzer, 1999; Pochapsky and Pochapsky, 2013) and Electron microscopy (Russell et al., 2004). Each of these methods employ atomic observation providing detail at the level of the protein residues involved in the interaction.

Two hybrid experiments provide less reliability but are more suited for high-throughput screening (Fields and Sternglanz, 1994). Two hybrid methods involve tagging one potential protein pair with the DNA-binding domain of a fragmented transcription factor and the other with the activation domain. If these proteins are close in proximity they will bring the DNA-binding and activation domains close enough to result in a functional transcription unit. As such this method allows us to directly infer protein interactions between two protein partners. Two hybrid methods are one of the more commonly used for protein interaction evidence in STRING’s database (von Mering et al., 2005).

Finally, multi-protein complexes are commonly detected using methods such as immunoprecipitation. Co-immunoprecipitation is used to experimentally infer suspected associations by isolating an individual protein using a suitable antibody and identifying any proteins that bind to it as an interactive partner. This type of experiment does not provide detail of the interaction at a chemical level or reveal which proteins are in direct contact . They do however give information as to which proteins are found in a complex at a given time and are suitable for high-throughput screens.

There are a number of additional physical interaction screening methods used as evidence in the primary sources used in STRING's database including Bimolecular fluorescence complementation, Affinity electrophoresis (Schou and Heegaard, 2006) , Phage display (Sidhu et al., 2003) and Proximity ligation assays (Söderberg et al., 2008).

1.5.2 Challenges associated with PPI data

Despite this range of methods and the continued work on the human PPI network there are still improvements to be made that might in turn improve models that use this data.

The primary issue with PPI data is its incompleteness (Mosca et al., 2013; Rolland et al., 2014). It is estimated that the human interactome hosts somewhere near 20,000 proteins that in turn share approximately 650,000 interactions (Amaral, 2008). Theofilatos et al. (Theofilatos et al., 2014) counted 210,000 catalogued protein interactions of various confidence and other studies of the time suggested that just 30,000 interactions of high confidence levels were publicly available, a small fraction of the estimated number of real interactions (De Las Rivas and Fontanillo, 2010; Rolland et al., 2014). In the studies presented here we use ~60,000 high confidence interactions.

As well as this incompleteness current protein-protein interaction data is generally non-directional so we do not know which protein in a pair is driving the interaction. Furthermore each interaction is generally weighted with just a confidence score and so we do not have information on the magnitude of an interaction or whether the interaction is inhibitory or excitatory in nature.

1.6 Network analysis

In the studies presented here we make use of physical interaction data and genetic interaction data to build graphical models and visualise our results (Jonsson and Bates, 2006; Manco et al., 2019)..

In these models each node represents a protein (interchangeable with the associated gene) and each edge represents an interaction between two nodes (Pavlopoulos et al., 2011). As discussed above the interactions in these PPI network representations do not have an intrinsic directionality nor do they have any weighting other than a confidence score which is not used in our models.

1.6.1 Topological features

A number of features can be extracted azfrom a network model of a biological system (Figure 1.5) (Almaas, 2007). We have broadly categorised these features as either pair-wise or node-wise depending on whether the feature is calculated based on the relationship between two nodes or associated with just a single node and its relationship to the rest of the network.

Commonly used examples of pair-wise topological features include:

Shortest path; which counts the minimal number of connected vertices that create a path between the source and target node.

Adjacency; A logical feature that states whether a source and target node are connected via an edge.

Mutual neighbours; A count of how many first neighbours a target and source node share.

Adhesion; The minimum number of edges that would have to be severed to result in two separate sub-graphs separating the source and target nodes.

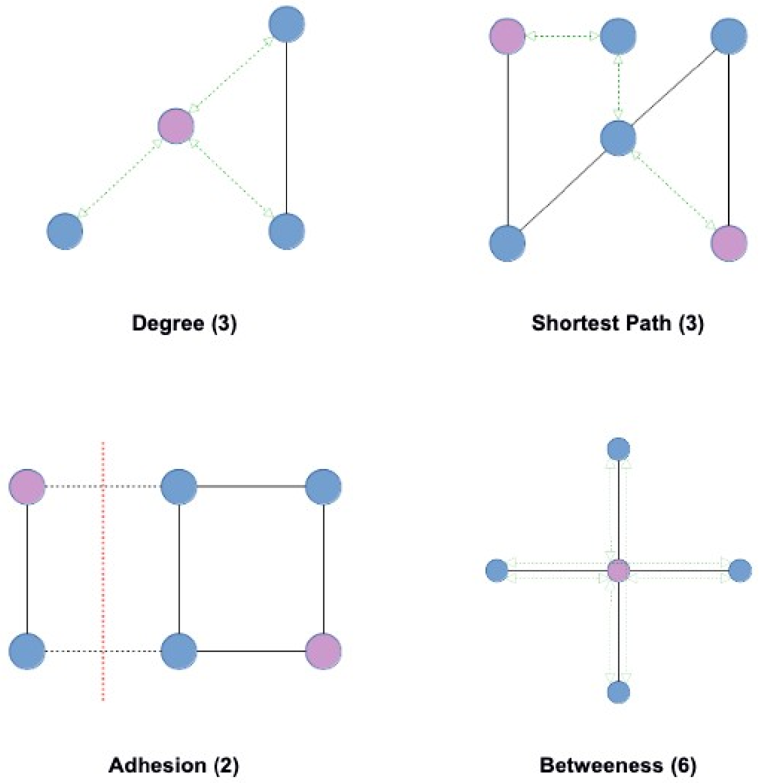
Some examples of node-wise features include:

Degree; The number of edges coming in to or out of the node.

Betweenness; The number of shortest paths in the entire graph that pass through the node.

Closeness; The number of steps required to reach all other nodes from a given node.

Eccentricity; The shortest path distance from the node farthest from the given node.



Degree (3)

Shortest Path (3)

Adhesion (2)

Betweeness (6)

Figure 1.5 Examples of topology features including degree, shortest path, adhesion and betweeness. The numbers in parentheses are the feature values as shown in the example, for example the pink node in the first graph example has a degree of three.

Many other topological features are introduced in the following studies as well as network features derived from random-walk simulations on the network and additional data sets such as gene ontology data.

1.7 Project Aims

In the following work I explore range of methods developed to identify vulnerabilities in cancer cells, with the ultimate aim of guiding the discovery of targeted and personalised cancer therapies. I describe computational systems approaches for predicting human synthetic lethal interactions by identifying and exploiting conserved patterns in protein-protein interaction network topology and the online database designed to publish the predictions from these classifiers.

I used similar network features to predict gene dependencies, otherwise known as acquired essential genes, in specific cancer cell lines by applying novel PPI network modelling and I analyse copy number variance and ploidy in kidney cancers to identify factors that may be driving this genomic instability.

Chapter 2 - Computational Approaches to Identify Genetic Interactions for Cancer Therapeutics

In this chapter I present a review of genetic interactions and SSL. The development of improved cancer therapies is frequently cited as an urgent unmet medical need. Here I discuss some of the shortcomings of traditional cancer therapies and how genetic interactions are being therapeutically exploited to identify novel targeted treatments for cancer that mitigate some of these shortcomings. I discuss the current methodologies that use ‘big data’ to identify genetic interactions, in particular focusing on synthetic sickness lethality (SSL) and synthetic dosage lethality (SDL). I describe the experimental and computational approaches undertaken both in humans and model organisms to identify these interactions. Finally I discuss some of the licensed drugs, the inhibitors in clinical trials and compounds under development, that are targeting SSLs and SDLs for the treatment of cancer. This chapter provided an in-depth literature review and background for the proceeding chapters.

Chapter 3 - Predicting synthetic lethal interactions using conserved patterns in protein interaction networks

In response to a need for improved treatments, a number of promising novel targeted cancer therapies are being developed that exploit human synthetic lethal interactions. This is facilitating personalised medicine strategies in cancers where specific tumour suppressors have become inactivated. Mainly due to the constraints of the experimental procedures, relatively few human synthetic lethal interactions have been identified. In this chapter I describe SLant (Synthetic Lethal analysis via network topology), a computational systems approach to predicting human synthetic lethal interactions that works by identifying and exploiting conserved patterns in protein interaction network topology both within and across species. SLant out-performs previous attempts to classify human SSL interactions and experimental validation of the models predictions suggests it may provide useful guidance for future SSL screenings and ultimately aid targeted cancer therapy development.

Chapter 4 - Slorth: Validated and predicted synthetic lethal gene pairs with associated drug, disease and orthology data

Chapter 4 presents Slorth (Synthetic Lethality and ORTHology), a database that enables the identification and analysis of synthetically lethal interactions (SSL) both in humans and in model organisms. The database documents 331,308 experimentally determined genetic interactions and 852,609 high quality synthetic lethal predictions obtained using the SLant algorithm. Powerful browsing and search functionality enables easy identification of putative SSL gene pairs which are integrated with cancer, drug, pathways and orthologue information highlighting those interactions that could be exploited therapeutically. Clear visualisation tools enable exploration of the wider network around the genetic interactions.

Chapter 5 - Biological network topology features predict gene dependencies in cancer cell lines

In Chapter 5 I investigate computational approaches that enable users to identify genes that have become essential in individual cancer cell lines. Using recently published experimental cancer cell line gene essentiality data, human protein-protein interaction (PPI) network data and individual cell-line genomic alteration data I build a range of machine learning classification models to predict cell line specific acquired essential genes. Genetic alterations found in each individual cell line were modelled by removing protein nodes to reflect loss of function mutations and changing the weights of edges in each PPI to reflect gain of function mutations and gene expression changes.

I found that PPI networks can be used to successfully classify human cell line specific acquired essential genes within individual cell lines and between cell lines, even across tissue types with AUC ROC scores of between 0.75 and 0.85. My novel perturbed PPI network models further improved prediction power compared to the base PPI model and are shown to be more sensitive to genes on which the cell becomes dependent as a result of other changes. These improvements offer opportunities for personalised therapy with each individual’s cancer cell dependencies presenting a potential tailored drug target.

The overriding motivation for predicting cancer cell line specific acquired essential genes is to provide a low-cost approach to identifying personalised cancer drug targets without the cost of exhaustive loss of function screening.

Chapter 6 - Defining signatures of arm-wise copy number change and their associated drivers in kidney cancers

In Chapter 6, using pan-cancer data from The Cancer Genome Atlas (TCGA), I investigate how patterns in copy number alterations in cancer cells vary both by tissue type and as a function of genetic alteration. I find that patterns in both chromosomal ploidy and individual arm copy number are dependent on tumour type. I highlight for example, the significant losses in chromosome arm 3p and the gain of ploidy in 5q in kidney clear cell renal cell carcinoma tissue samples. Using signatures derived from non-negative factorisation I also find gene mutations that are associated with particular patterns of ploidy change.

Finally, utilising a set of machine learning classifiers I successfully predicted the presence of mutated genes in a sample using arm-wise copy number patterns as features. This demonstrates that mutations in specific genes are correlated and may lead to specific patterns of ploidy loss and gain across chromosome arms. Using these same classifiers, I highlight which arms are most predictive of commonly mutated genes in kidney renal clear cell carcinoma (KIRC).

Chapter 7 - Discussion

In this chapter I discuss how the work presented in this thesis has achieved the goal of furthering the field of personalised cancer treatment and new target discovery. I highlight the novel aspects of the work that have led to valuable new methods, insights and predictions and I discuss some of the challenges faced and limitations that have been experienced.

2 - Computational Approaches to Identify Genetic Interactions for Cancer Therapeutics

2.1 Introduction

Cancer is a genetic disease that develops as a result of a number of mutational events caused by endogenous and exogenous processes. The resulting mutations enable a cancer cell to gain a selective advantage over healthy cells, often resulting in uncontrolled proliferation and ultimately metastasis of a cancer (Douglas Hanahan, 2000; Hanahan and Weinberg, 2011). Cancer therapies must by necessity attack the aberrant cells once a tumour is discovered. However, established chemotherapy regimes often affect targets shared by normal and cancer cells and often kill “healthy” but rapidly dividing cells. This leads to significant damage in unintended targets resulting in the trademark side-effects of cancer therapy; gastrointestinal upset and hair-loss (Coates et al., 1983) .

The therapeutic index (TI) is a comparison of the amount of a therapeutic agent that causes the therapeutic effect to the amount that causes toxicity. Standard chemotherapies often have a low TI due to the challenge presented by selectively targeting cancer cells whilst sparing normal cells (Muller and Milton, 2012). Furthermore, due to cancer cells' predisposition to acquire mutations, a drug that seems effective at the outset of therapy may well be rendered ineffective if even a single cell, and its resulting daughters, gain resistance to that compound (Holohan et al., 2013).

2.2 Targeted therapies

Of the 154 cancer drugs that are licensed by the FDA, 85 are new, targeted therapies often targeting the genes that directly drive cancer (Santos et al., 2016). These driver genes can be broadly classified as oncogenes or as tumour suppressors. When mutated, the protein products of oncogenes show an increase in activity, or a gain or change of function (GOF) that result in tumorigenesis. Conversely in tumour suppressors, mutations (or epigenetic silencing) result in the loss of function (LOF) of the protein product.

Many targeted anticancer drugs work by directly inhibiting activated oncogenes, particularly proteins that are nuclear receptors or those that contain protein kinase domains (Iorio et al., 2016; Nguyen et al., 2017; Shawver et al., 2002). Dabrafenib, which has been approved for the treatment of late-stage melanoma, targets the constitutively activated kinase oncogene BRAF V600E. Whilst gefitinib and erlotinib licensed for the treatment of lung cancer targets the EGFR tyrosine kinase (Lindeman et al., 2013; Shepherd et al., 2005; Stinchcombe and Socinski, 2008; Thatcher et al., 2005). A substantively different approach is needed to provide therapies aimed at controlling the damage done by inactivated tumour suppressor genes. It is not usually feasible to repair the protein products of these genes particularly if they are inactivated by truncation, although there are on going attempts to reactivate or restore function to a small subset of p53 missense mutant proteins. These attempts to develop drugs to reactivate TP53 have led way to another class of therapy, anti-inhibition. Inhibitors of MDM2, a negative regulator of p53, have shown some promise in restoring function in the P53 pathway including apoptosis which can lead to tumour regression. A number of compounds related to nutlin-3a, a class of small molecule MDM2 inhibitors, are currently in phase I or II trials (Burgess et al., 2016; Hoe et al., 2014).

2.3 Genetic interactions

Genetic interactions are when mutations in two genes (or alternatively the loss of two genes) produces a phenotype that is enhanced in comparison to each mutation’s (or gene loss) individual impact. This phenomenon can reveal functional relationships between genes and pathways (Krause and Gray, 2009). Two types of genetic interaction are of particular interest in the field of cancer drug development; synthetic lethality (SSL) sometimes termed “synthetic sick lethality, and synthetic dosage lethality (SDL) (described below). Here we describe how the identification of these genetic interactions is being used to guide therapeutic strategies for the treatment of cancer.

A natural redundancy of function in our cells allows for a number of otherwise essential pathways to be disrupted by mutations whilst allowing the cell to remain viable. In some cases these disruptions can lead to impaired function of important cell maintenance or regulatory pathways leading to an increased occurrence of mutations or increased proliferation. These mutations are often found in tumour samples as they can often confer an increased fitness over normal cells.

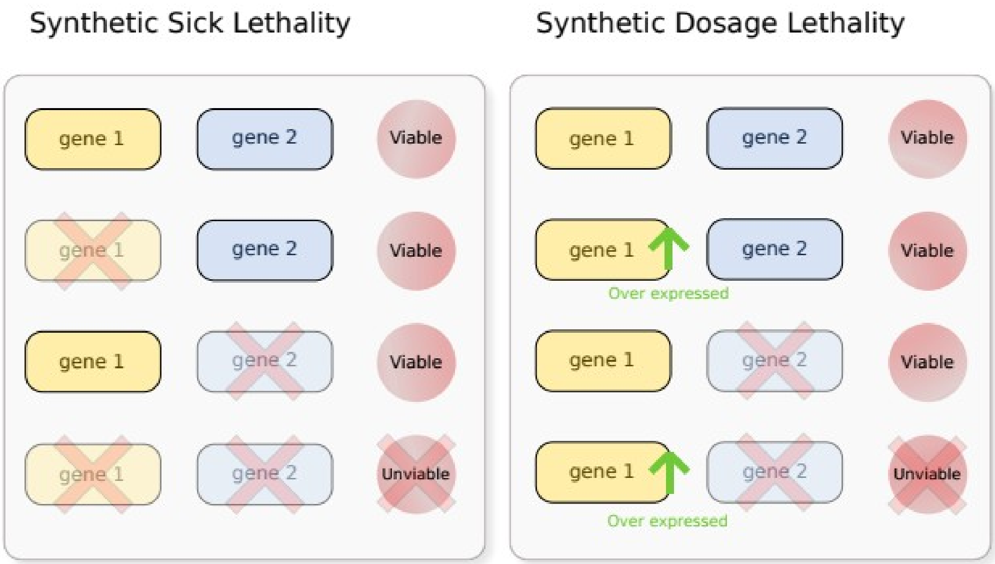


Figure 2.1: Schematic illustration of synthetic sickness lethality and synthetic dosage lethality. In the case of SSL gene pair the cell remains viable when either gene is individually deleteriously altered, when both genes are altered concurrently the cell loses viability. In the case of SDL the cell loses viability when one gene is deleteriously altered while the other is up-regulated.

This redundancy gives rise to the possibility of SSL, where individuals in a pair (or more) of genes can be disrupted without affecting cell viability whilst disruptions in both genes causes cell sickness or death. Two genes are said to be synthetic lethal when a concurrent deleterious mutations or complete deletion of both leads to the death of the host cell whilst a mutation or deletion in either alone leaves the cell viable (Hartwell et al., 1997). SSL is occasionally termed 'synthetic sensitive lethality'' or 'synthetic sick lethality', these terms are commonly used interchangeably in the literature.

Synthetic dosage lethality (SDL) interactions occur when over-expression of gene A is lethal when gene B has a loss of function (Figure 2.1).

2.4 Using genetic interactions as a therapeutic strategy

To exploit genetic interactions therapeutically, the genetic defects in an affected pathway must be combined with a pharmacologically induced defect in a compensating pathway. Synthetic lethality is well suited for targeting deactivated tumour suppressors (Megchelenbrink et al., 2015). SSL causes cell death as a result of one gene being genetically inactivated by mutation (LOF, the tumour suppressor) and another being inactivated by a drug target. While synthetic dosage lethal interaction can be used for targeting cancer cells with over-expressed oncogenes (Megchelenbrink et al., 2015). SDL causes cell death as a result of one gene being genetically activated (GOF, the oncogene) and another being inactivated (LOF, the drug target).

Exploiting SSL/SDL pairs as drug targets may provide significantly improved therapeutic indices of our drugs compared to standard chemotherapies by selecting exclusively for cancer cells harbouring mutations in pathways that make part of a synthetic lethal pair (McLornan et al., 2014).

To compound the problem some mutations that occur later in the evolution of cancer may be tolerated due to earlier mutations. This network of interactions may prove extremely complicated though we may find that pathways activated early in tumour progression are likely to make better targets for analysis (Kaelin Jr and Kaelin, 2005).

2.4.1 Methods that identify genetic interactions

Although there are some insights into where SSL interactions are likely to occur, for example Matteo et al. (D'Antonio et al., 2013) found an enrichment of SSL interactions between recessive cancer genes and their functional paralogues, identifying SSL interactions is a hard problem. Due to experimental limitations not many SSL interactions in humans have been published, but more is known about those in model organisms.

Approximately 20% of genes in *Saccharomyces cerevisiae* (*S.cerevisiae*) are essential (Tong et al., 2004) which leaves the others to have the potential to exhibit genetic interactions. Systematic double-knockout screens on large subsets of genes is *S. cerevisiae* and *Caenorhabditis elegans* (*C. elegans*) suggest that, on average, 0.5% of tested gene pairs are synthetic sick or synthetic lethal, and that many SSL interactions involve more than two genes. The result is a combinatorial problem for the traditional screening of all possible interactions. This and our limited data on these molecular networks prevents easy, reliable systematic prediction of SSL interactions (Chipman and Singh, 2009).

2.4.1.1 Experimental approaches to identify genetic interactions

Synthetic Lethality was first described in 1922 by Calvin Bridges in a study crossing Drosophila melanogaster and later named by Theodore Dobzhansky, this time crossing Drosophila pseudoobscura, in 1946 (Nijman, 2011). Similar to these early experiments contemporary SSL studies have classically focused on crossing eukaryotic model organisms with increasing sophisticated techniques allowing researchers to mutate and mate hybrid genomes and screen using gene silencing techniques such as RNA interference (RNAi).

More recently high throughput approaches to finding genetic interactions in model organisms have been developed based broadly around three distinct platforms; synthetic genetic array (SGA) (Tong et al., 2004), diploid based synthetic analysis on microarrays (dSLAM) (Pan et al., 2007) and epistatic miniarray profiles (E-MAP) (Collins et al., 2010). Tong et al.'s SGA assay in S. cerevisiae uses a yeast strain with a single disabled gene and mates it with an array of yeast strains each with an individual deletion resulting in approximately 4,700 mutation pairs with varying viability. These techniques were further refined in Ooi et al.'s SLAM (Ooi et al., 2003) and again in Pan et al.'s (Pan et al., 2007) dSLAM. SLAM generates ordered arrays of double yeast knockout mutant sets (YKO) where the query mutation is introduced by integrative transformation rather than mating, and a microarray readout is used to produce a ranked list of candidate genetic interaction genes. In dSLAM, a pool of all heterozygous deletion diploids is transformed en masse with a single query gene disruption construct after which single- and double-mutant haploid pools are derived by sporulation and differential selection. These techniques have been extended from S. cerevisiae to Saccharomyces pombe (S. pombe), C. elegans and Escherichia coli (E. coli) significantly increasing the quantity and quality of genetic data available. Collins et al.'s (Collins et al., 2010) E-MAP performs SGA on a subset of genes selected specifically from a pathway or functional grouping.

Although it was at one time hypothesised that SSL pairs could be conserved across species if both species shared orthologues for those respective genes (Wu et al., 2013), it has since been found that in many cases these SSL interactions are not conserved between lower eukaryotes and their human orthologous equivalents. As such, though SSL data for model organisms can teach us a lot about gene function and pathway interaction, our search has been extended to human cell lines and those of phylogenetically similar organisms.

Classically SSL discovery in humans has been a ‘hypothesis driven’ process of predicting SSL interactions based on proven associations, often related to loss of particular cell cycle checkpoints or pathways related to those of known tumour suppressors, and subsequent clinical trials. However, with the increasing availability of genetically modified human cell lines and high throughput genetic screening methods that combine RNAi screens with libraries of small molecule inhibitors, an increasing number of human SSLs are being identified (Barbie et al., 2009; Berns et al., 2004; Iorns et al., 2007; Luo et al., 2009a; Scholl et al., 2009) .

2.4.2 Computational techniques used to predict genetic interactions

A systematic approach to inferring genetic interactions has become increasingly popular in the past decade. The ever-growing amount of screening data available has paved the way for more sophisticated computational techniques employing statistical and machine learning (Table 2.1).

Data Type	**Database**	**URL**	**Reference**
Protein interactions	STRING	http://string-db.org/	(von Mering et al., 2005)
Gene expression	Expression Atlas	https://www.ebi.ac.uk/gxa/home	(Petryszak et al., 2016)
	Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo/	(Edgar, 2002)
Gene coexpression	CoxpresDB	http://coxpresdb.jp/	(Okamura et al., 2015)
Gene ontology data	Gene Ontology Consortium	http://www.geneontology.org/	(Ashburner et al., 2000)
Somatic mutations	Cosmic	http://cancer.sanger.ac.uk/cosmic	(Bamford et al., 2004)
Homology	Ensembl - comparative genomics	http://www.ensembl.org/info/genome/compara/index.html	(Herrero et al., 2016)
Cellular phosphorylation	Networkin	http://networkin.info/	(Linding et al., 2008)
Integrative – multiplatform data	The Cancer Genome Atlas (TCGA)	cancergenome.nih.gov	(Tomczak et al., 2015)
	The International	icgc.org	(Zhang et al., 2011)

	Cancer Genome Consortium (ICGC)		
	cBioPortal	www.cbioportal.org	(Cerami et al., 2012)
	MOKCa	strubiol.icr.ac.uk/extra/mokca	(Richardson et al., 2009)

Table 2.1. List of data sources that have been used to extract predictive features used for genetic interaction classification

These in silico models have proved significantly cheaper and faster to implement compared to traditional screening methods and have demonstrated high levels of accuracy when predicting genetic interactions as discussed below.

These studies can be broadly classed by the type of parameter, or feature in the context of machine learning, used to train their model. The most prevalent parameter types include biological network data, gene ontology and expression level data, and orthology or evolutionary data though a number of studies use a combination of these data. Whilst this review has more emphasis on human SSL interactions we do discuss a number of studies focused on model organisms as much work on human genetic interactions has foundations in early work on lower eukaryotes.

Table 2.2 summarises the SSL training datas sources commonly used in the studies below.

Source	**Species**	**Number of SSL pairs**	**URL**	**Reference**
Biogrid	H. sapiens	503	https://thebiogrid.org	(Stark, 2006)
	S. cerevisiae	92,738		
	D. melanogaster	3		
	C. elegans	1,237		
	S. pombe	36,353		
SynLethDB	H. sapiens	19,952	http://histone.sce.ntu.edu.sg/SynLethDB	(Guo et al., 2015)
	S. cerevisiae	13,421		
	D. melanogaster	423		
	M. musculus	366		
	C. elegans	107		

The Cellmap	S. cerevisiae	1,198	http://thecellmap.org	(Dallago et al., 2018)
Flybase	D. melanogaster	9,661	http://flybase.org/	(Gelbart et al., 1996)
Other studies	S. cerevisiae	100		(Collins et al., 2007)
	C. elegans	1,246		(Byrne et al., 2007)

Table 2.2. A summary of the SSL training datas sources commonly used in the studies below.

2.4.2.1 Biological network data approaches

A number of studies have employed a systems approach to predict SSL interactions using network parameters extracted from biological network data. These biological networks include data such as physical interactions and co-expression (Figure 2.2).

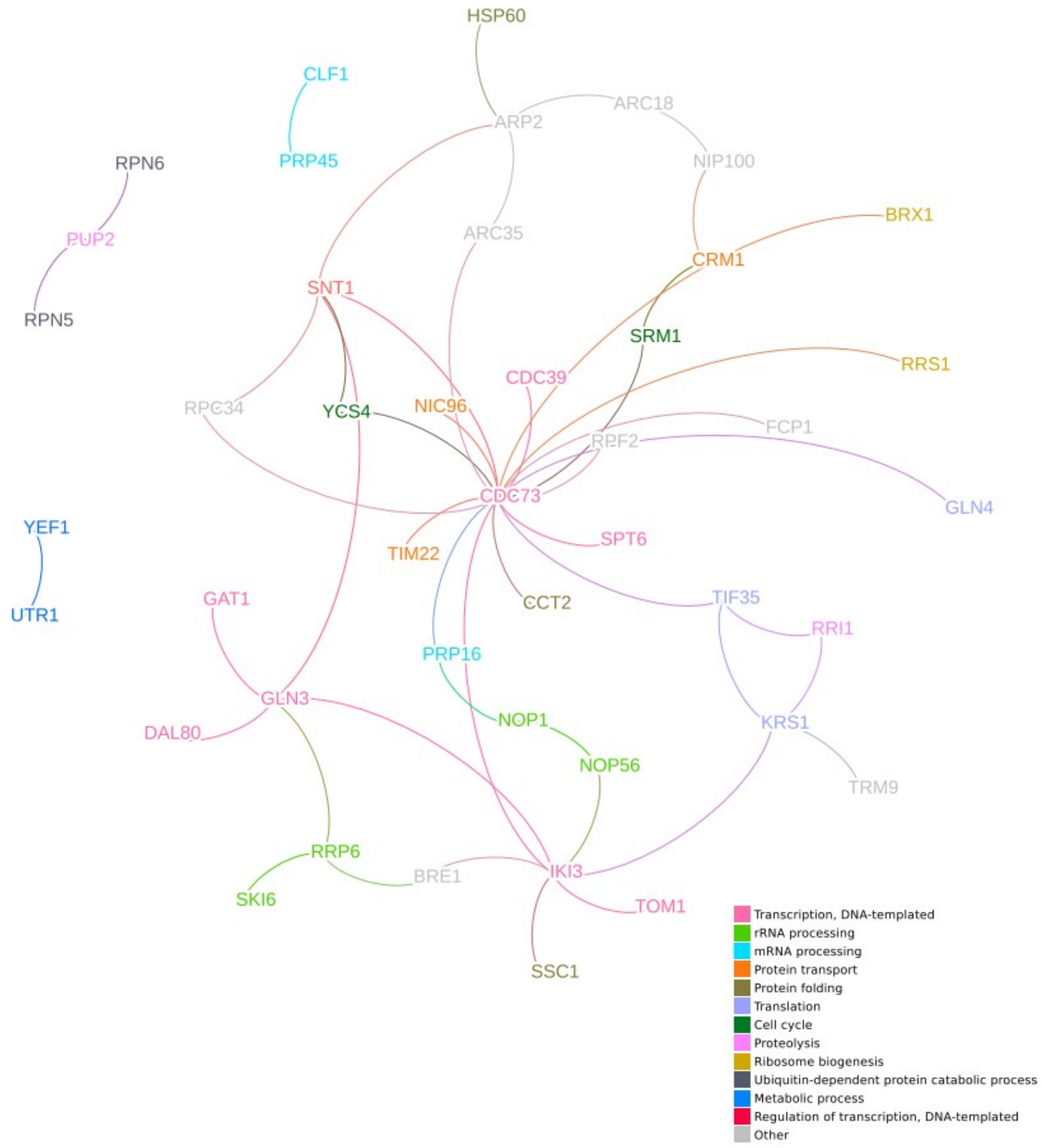


Figure 2.2: An example of biological network data - A genetic interaction network of S. cerevisiae DDR genes coloured by GO terms. Genetic interaction data collected from BioGRID and filtered for orthologues of known human DDR genes. GO terms sourced from Gene Ontology Consortium and filtered for functional terms only, the most popular overall GO term was chosen for genes with multiple annotations.

Early attempts to predict genetic interactions such as Wong et al. (Wong et al., 2004b) utilised decision tree classifiers trained on biological networks data including a number of topological network features from protein to protein interaction graphs, gene co-occurrence data and mRNA co-expression data. This study predicted 740 SSL interactions in 2,356 possible pairs in *S. cerevisiae* with a success rate of 0.31, a vast improvement on the 0.0056 success rates achieved by previous unguided approaches. This approach was extended by Zhong et al. (Zhong and Sternberg, 2006) to predict interactions in C. elegans, an organism with relatively less available genetic interaction data, through orthology. By training a model using features from the relatively large datasets from yeast and fly models this study was able to predict interactions across species using logistic regression. Further attempts to predict genetic interactions in *S. cerevisiae* using biological networks followed as Paladugu et al. (Paladugu et al., 2008) extracted multiple features from protein–protein interaction networks, which were applied to a SVM classifier to predict new SSL interactions with sensitivity and and specificity exceeding 85%.

By employing random walks and decision tree classifiers on biological networks including protein-protein interactions, GO interactions and existing known genetic interaction data Chipman et al. (Chipman and Singh, 2009) were able to predict synthetic lethal interactions at a true positive rate of 95 percent against a false positive rate of 10 percent in *S. cerevisiae*. And a true positive rate of 95 against a false positive rate of 7 percent in C. elegans. They noted that including experimentally validated non-interactions into training data significantly improved results for both organisms.

While the majority of preceding studies focused on supervised learning You et al. (You et al., 2010) performed semi-supervised learning on both the functional and topological properties of a functional gene network, this network was a result of the integration of protein to protein interaction data along with protein complex and gene expression data and resulted in a maximum accuracy of true positive rate of 92% against a false positive rate of 9%.

Attempts to predict SSL interactions using expression data as a primary training parameter led to Bandyopadhyay et al.'s (Bandyopadhyay et al., 2011) SSLPred used regression on training data that mapped expression levels between gene with single deletion mutations to their corresponding SGA entries to predict SSL interactions. Again using expression level data but this time to predict SSL in the context of somatic mutations in TP53, Wang et al. (Wang and Simon, 2013) selected a number of genes which encoded kinases that exhibited significantly higher expression in tumours with functional p53 somatic mutations than in tumours without. These pairs were treated as potentially druggable synthetic lethal pairings for TP53 and many were confirmed via previous RNAi screenings.

To further improve results through an ensemble machine learning model Zheng et al. (Wu et al., 2014) developed MetaSL, a model boasting 17 features (11 similarity based features and 6 lethality based features) which was applied to 8 classifiers; random forest, J48 (a type of decision tree), Bayesian logistic regression, Bayesian network, PART (a rule-based classifier), RBFNetwork, bagging (bootstrap aggregating), and classification via regression. The predictions from these classifiers were aggregated yielding ROC AUC scores of 87.1% on yeast data. In another novel approach Zhang et al. (Zhang et al., 2016) modelled influence propagation in signalling pathways employing values of phosphorylation levels between signalling proteins in a similar way to that of studies modelling influence across social media platforms. A number of reliable, novel SSL pairs were predicted along with known interactions using this method.

Building on Zhong et al.'s attempt to predict SSL interactions using training data across species Jacunski et al. (Jacunski et al., 2015) developed SINaTRA (Species-Independent TRAnslation) to compare orthologous gene pairs between S. cerevisiae and S. pombe along with their respective physical interaction data (including 4 pairwise parameters and 20 ontological features) to calculate what was termed connectivity homology to improve prediction of orthologous interactions. This approach achieved a reported ROC AUC score of 0.86 when predicting SSL interactions between the two studied yeast species. The model trained on yeast data was applied to predict 1,309 human SSL pairs with a reported false positive rate of 0.36%

2.4.2.2 Evolutionary approaches

Although genetic interactions are not reliably conserved between species, with as little as ~23% of the interactions conserved between S. cerevisiae and S. pombe (Wang and Simon, 2013), and even less conservation between lower and higher eukaryotes, a number of research groups have managed to use orthological and evolutionary data to infer SSL interactions in humans.

By integrating phylogenetic analysis and data including interactions from BioGRID for interactions, homology from Ensembl and NCBI and GO attributes from Gene Ontology, Conde-Pueyo et al. (Conde-Pueyo et al., 2009) reconstructed a phylogenetically-inferred SSL gene network for humans. The culmination of this study was to identify a number of genes related to cancer cells (ATM, NF1, FBXW7, MSH2, BUB1, ERCC2, BLM and

MSH6) likely to be in therapeutically viable SSL interactions.

In a set of related studies researchers attempted to describe the mechanics of genetic interactions as a function of evolution and how these mechanics are conserved across species. VanderSluis et al. (VanderSluis et al., 2010) attempted to elucidate the evolutionary trajectories of duplicate genes through genetic interaction data and, as expected, found significant enrichment of genetic interactions between duplicate genes. Koch et al. (Koch et al., 2012) went on to describe how the rules governing genetic interactions are conserved across species. Using S. cerevisiae as a model they predicted the genetic interaction degree (i.e. how well connected a gene is in a genetic interaction network, or how many other genes a particular gene interacts with) for a number of S. pombe genes with high accuracy. Conserved features used to predict this degree of interaction included a quantitative measurement of single mutant fitness defects of the gene, multi-functionality, degree in a protein to protein interaction network and expression variation of the gene.

Lu et al. (Lu et al., 2013) also inferred human SSL pairs in human protein complexes by exploiting the evolutionary history of genes in parallel converging pathways in metabolism. This approach predicted around 250 novel SSL interactions 36 of which had a least one cancer related gene.

2.4.2.3 Integrative data approaches

As well as network base systems biology approaches and evolutionary methods a number of studies have also utilised the wealth of functional data such as mutation and copy number profiles, co-expression and functional relationships such as pathway data to predict synthetic lethal interactions.

In an early attempt to use a branch of natural language processing alongside biological data Pesquita et al. (Pesquita et al., 2009) focused on the semantic similarity of GO terms, codes used as a proxy for functional pathways, to successfully compare the functionality of two genes. This technique was later used by Kamath et al. (Kamath et al., 2003) as a method for predicting genetic interactions in a number of model organisms. In 2011 Li et al. (Lu et al., 2013) attempted to use an expectation-maximization algorithm on domain genetic interaction data to predict SSL interactions. It was reported that this approach was able to predict 17 novel SSL interaction in *S. cerevisiae* with probability > 0.9. Including the MYO4 – DYN1 pair with a probability of 0.9895. These interactions were further used to predict a number of compensatory pathways.

A number of algorithms have also been introduced that predict pairs of genes that would potentially exhibit genetic interactions using cancer data directly by identifying and scoring of sets of genetic alterations where the mutations are mutually exclusive, I.e. if gene 1 and gene 2 are synthetically lethal there should be no samples where both these genes are switched off. Initially these models used scoring regimes to prioritise mutual exclusivity with no basis in statistics. However, the approaches have gradually been refined to improve statistical scoring of the results and to integrate different methods of identifying whether or not a gene has essentially been switched off. These include Recurrent Mutually Exclusive aberrations (RME) (Miller et al., 2011) that uses mutation and copy number variation (CNV) data from 145 glioblastoma samples from The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015), and CoMet (Leiserson et al., 2015) that used mutation and CNV data from five TCGA studies. The more sophisticated of these, CoMet, looks at small groups of mutually exclusive genes, using a hypergeometric distribution to work out the probability of getting at least as unexpected a result as that seen. Using similar methods Srihari et al. (Srihari et al., 2015) analysed mutual exclusivity in copy number and gene expression data from four cancers to identify 718 genes that potentially share a SSL interaction with at least 1 of 6 DDR genes related to those cancers.

Another approach is the DAISY workflow (Ryan et al., 2014) which uses three inference procedures to identify both SSL and SDL pairs using data from cell lines as well as from clinical samples; somatic copy number variation and mutation profiles, shRNA-based functional examination and pairwise gene co-expression. DAISY was applied to VHL, PARP1, MHS2 and KRAS and achieved an AUC score of 0.779 demonstrating a strong propensity (p-value < 1x104) for predicting SSL pairs.

2.4.3 Cancer therapies that exploit genetic interactions

Many studies screening for genetic interactions have naturally focused on known cancer driver genes, specifically tumour suppressors, as promising targets for developing cancer therapies. Genetic interactions have been found in a wide range of cell pathways including cell cycle progression and apoptosis pathways.

Tumour suppressors that make part of DNA damage response pathways are prime candidates for synthetic lethal drug targets (Pearl et al., 2015). BRCA1 and BRCA2, both important in repair of double strand breaks, have been shown to share a synthetic lethal relationship with PARP (poly(ADP-ribose) polymerase), an important gene in single strand break repair. Cells deficient in either BRCA gene are extremely sensitive to PARP inhibitors presenting therapeutic opportunities (Bryant et al., 2005). Further studies systematically screening genes for sensitivity to PARP inhibitors identified a number of kinases whose inhibition strongly sensitised the host cell to PARP inhibitor, including cyclin-dependent kinase 5 (CDK5), MAPK12, PLK3, PNKP, STK22c and STK3 (Turner et al., 2008). There are number of PARP inhibitors at different phases of trials, a notable example being olaparib (Lynparza™, Astrazeneca) which has already been approved by both the European commission and the US food and drug administration for the treatment of patients with advanced ovarian cancer paired with BRCA mutations (Liu et al., 2014; Tangutoori et al., 2015). As well as a treatment for ovarian cancer patients Mateo et al. (Mateo et al., 2015) conducted trials for olaparib as a potential therapy for prostate cancer patients identified as having homozygous deletions, deleterious mutations or both in DNA-repair genes including BRCA1 or BRCA2, ATM, Fanconi's anemia genes, and CHEK2. Of the patients available for evaluation 88% responded to olaparib including all patients with BRCA loss leading to the conclusion that the drug led to a high response rate in prostate cancer patients with DNA- repair defects who were no longer responding to standard treatments. Recent PARP inhibitor based therapies include rucaparib which has also received FDA approval for patients with advanced ovarian cancer who suffer germline or somatic BRCA1 or BRCA2 mutations (Shirley, 2019; Syed, 2017) and Talazoparib which is showing promise in early trials for early-stage breast cancer patients with BRCA mutations even before any chemotherapy or surgery with all patients exhibiting a reduction in tumor size after 2 months (Litton et al., 2017).

Although PARP inhibition has enjoyed a lot of attention in drug discovery research into other other genetic interactions has also resulted in some potential therapies that target other pathways. MTH1 inhibition for example has been validated as an anti-cancer target. Although MTH1, a protein which sanitizes oxidized dNTP pools to avoid incorporation of damaged bases during DNA replication, is non-essential in normal cells, cancer cells require MTH1 activity. Without MTH1 cancer cells risk incorporation of oxidized dNTPs which leads to DNA damage and cell death. Gad et al. Describe two small molecules, TH287 and TH588, in the nudix hydrolase family that selectively inhibit MTH1 protein in cells (Gad et al., 2014).

Other published SSL with possible therapeutic potential include the SSL interaction between TP53 and the PI5P4K gene family where the PI5P4K kinases are essential for growth in the absence of p53 (Emerling et al., 2014), ARID1A, a chromatin remodeller with a high mutation rate across many cancer types shared a SSL interaction with the EZH2 methyltransferase in ARID1A-mutated ovarian cancer cells (Bitler et al., 2015) and ENO2 which selectively inhibits viability of ENO1-deleted glioblastoma cells (Muller et al., 2012). PTEN, a gene associated with genomic stability, and APE1, important in DNA base excision repair, have been shown to share a SSL relationship with treatment of APE1 inhibitors in PTEN-deficient cells resulting in the induction of apoptosis (Abbotts et al., 2014). ATR, an important DNA damage response gene has also been identified as potential synthetic lethal pair of ARID1A and a number of ATR inhibitors are in phase I trials as a potential therapy for ARID1A deficient tumors (Karnitz and Zou, 2015; Williamson et al., 2016).

While much work on genetic interactions as therapy targets has traditionally focused on SSL interactions research has also been conducted into the SDL interactions of several potent oncogenes such as MYC and KRAS (Workman et al., 2013). Members of the RAS superfamily are some of the most commonly activated cancer drivers (Pylayeva-Gupta et al., 2011) and showed some promise in early SDL research. These studies described a number of potential SDL pairs including an interaction between K-RAS and CDK4 which offers potential opportunities in non-small cell lung carcinoma therapy (Puyol et al., 2010). Another systematic study of the RAS superfamily found a number of interactions with genes related to the cells mitotic functions including PLK1 (Luo et al., 2009a). Despite this early promise other therapies related to RAS such as the direct targeting of the RAS protein and immune checkpoint blockade have proved more effective and no promising new therapeutic approaches related to SDL interactions have been discovered to date (Downward, 2015; Pylayeva-Gupta et al., 2011). Improved screening through CRISPR-cas9-based techniques may provide further potential SDL interactors for mutant RAS genes in future studies (Papke and Der, 2017).

Other genes with potential SDL interactions with KRAS include CDK1, part of the Cyclin-dependent kinase family with CDK4 which is mentioned above (Costa-Cabral et al., 2016), TBK1, a serine / threonine kinase important in regulating inflammatory response (Barbie et al., 2009) and GATA2, essential in regulating transcription (Kumar et al., 2012).

2.5 Discussion

The performance of contemporary models used to predict SSL interactions is difficult to assess due to a lack of a gold standard source of human SSL pairs. This difficulty is compounded by the lack of a single extensive, curated repository of known human SSL pairs. Furthermore the actual number of potential human SSL pairs is so far unavailable proving another challenge when attempting to assess progress in the field.

In studies employing a CRISPR-Cas9-based screen of 18,166 human genes only 1,878 were essential, resulting in 16,288 non-essential genes (or as much as 90.8% of the whole genome) each potentially part of at least one genetic interaction. Despite this large number of potential synthetic lethal interactions only 503 human gene pairs are classed as synthetic lethal or negative genetic in the BioGRID, a current primary source for curated validated human SSL pairs. There are many more predicted synthetic lethal pairs documented in sources such as SynLethDB which collates 19,952 predicted pairs sourced from in-silico predictions from tools such as Daisy (which counts 5,824 SSL pairs), shRNA screening experiments and literature via text-mining though the reliability of many of these observations is very hard to quantify.

Of course the goal of listing and validating all possible synthetic lethal interactions may be neither possible or even valuable in the context of therapeutics, the majority of synthetic lethal pairs will likely not pertain to cancer genes or perhaps even genes with any therapeutic value. The majority of research has been focused around cancer related genes with many of the studies outlined above focus on a small subset of interactions focused around notable cancer drivers such as BLM (Conde-Pueyo et al., 2009), TP53 (Wang and Simon, 2013), VHL, PARP1 and MHS2 (Jerby-Arnon et al., 2014) amongst others. So far no meta-analysis has been completed on these disparate studies though this might be a good first step towards making SSL interaction data more coherent.

The majority of standard chemotherapies exhibit a very low therapeutic index (TI). In these therapies the level of treatment that is likely to cause toxicity in a patient is not significantly higher than the level that offers a therapeutic effect. To improve this therapeutic index and, a result, the quality of life and prognosis of cancer patients our goal must be to discover targets that can be drugged to selectively affect cancer cells whilst leaving normal cells unharmed. By exploring and exploiting vulnerabilities presented by genetic interactions (GI) and, more specifically, synthetic sensitive lethality (SSL) interactions in human cancer cells we may find ways to provide personalised care with both an increased therapeutic index and ultimately an improved prognosis for the cancer patient. While SSL interactions may present a unique opportunity in the fields of drug discovery and personalised cancer medicine the genome-wide identification of human SSL interactions comes with its own significant challenges. As well as the difficulty of propagating human cell lines for in-vitro screening the combinatorial nature of the problem means that around 200 million pairwise tests would be required to identify all possible pairs, an all but insurmountable experimental burden.

In response to these difficulties studies focussing on model organisms with far fewer genes and no ethical implications have resulted in the identification of a large quantity of SSL interactions. Unfortunately, based on these studies, it has been shown that SSL interactions are often not well conserved between species and even less so between higher and lower eukaryotes such as humans and yeast.

Though a number of unique human SSL interactions have been inferred using orthologous interactions many remain undiscovered and the search for SSL interactions opens to ever increasing quantities of multi-platform genomic data to develop a systematic approach for predicting potential SSL interactions utilising in-silico models. Ultimately this work will utilise the wealth of genomic, proteomic data available to explore the networks of both the humans interactome and those of our model organisms. From this we hope to understand how interactions are conserved and how we might better predict human SSL interactions given publicly available resources.

3 - Predicting synthetic lethal interactions using conserved patterns in protein interaction networks

3.1 Introduction

Despite sustained global efforts to develop effective therapies, cancer is now responsible for more than 15% of the world's annual deaths. There are over 12 million newly diagnosed cases per annum and this number continues to grow (Varmus and Kumar, 2013). Standard chemotherapy involves non-selective, cytotoxic agents that often have limited effectiveness and strong side-effects. Consequently, the current focus in oncology drug discovery has moved towards identifying targeted therapies that promise both improved efficacy and therapeutic selectivity (Yap and Workman, 2012).

The development of multi-platform genomic technologies has enabled the identification of many of the genes that drive cancer (Tomczak et al., 2015). These cancer driver genes can be broadly classified either as oncogenes or tumour suppressors. The protein product of an oncogene shows an increase in activity, or a change or gain of function when mutated, whereas mutations or epigenetic silencing in tumor suppressors result in an inactivation or loss of function (LOF) of the protein product (Baeissa et al., 2017a).

Targeted therapies that act on oncogenes often work by directly inhibiting the activated protein product. This strategy has been particularly successful for targeting nuclear receptor proteins or those that contain protein kinase domains (Iorio et al., 2016; Nguyen et al., 2017; Shawver et al., 2002). Unfortunately, it is not usually feasible to repair tumour suppressor genes or their protein products, particularly if they are inactivated by a truncation (Hoe et al., 2014). Instead an emerging strategy is to target tumour suppressors indirectly by exploiting synthetic lethal interactions.

Synthetic lethality (SSL) is a phenomenon whereby individual genes in a pair can be knocked-out without affecting cell viability, whilst disruptions in both genes concurrently cause cell death (Hartwell et al., 1997). Synthetic sensitive and synthetic sickness interactions are extensions of this concept where concurrent genetic interactions impair cellular fitness without necessarily killing the cell. Conversely, synthetic dosage lethality (SDL) interactions occur when over-expression of one gene, in combination with loss of function in another gene results in cell death. SSL and SDL interactions are both examples of negative genetic interactions. Negative genetic interactions are events where a deviation from the expected phenotype is observed when genetic mutations occur in more than one gene (Michaut and Bader, 2012).

To exploit SSL interactions therapeutically one gene, the tumour suppressor, is genetically inactivated by mutation while the protein product of the other is targeted and inactivated pharmacologically (Megchelenbrink et al., 2015). Synthetic dosage lethal interactions can be used for targeting cancer cells with over-expressed, undruggable oncogenes (Megchelenbrink et al., 2015). SDL causes cell death as a result of one gene being genetically activated (GOF, the oncogene) and another being inactivated (LOF, the drug target).

PARP inhibitors are the most developed therapies that exploit SSL interactions. The PARP inhibitor Olaparib (Lynparza™, Astrazeneca), has been approved for the treatment of patients with recurrent, platinum-sensitive, high-grade serous ovarian cancer with BRCA1 or BRCA2 mutations (Liu et al., 2014; Tangutoori et al., 2015). PARP1 (poly(ADP-ribose) polymerase) is an important component in DNA single strand break repair and has been shown to share a synthetic lethal relationship with both BRCA1 and BRCA2 (Aguilar-Quesada et al., 2007; Farmer et al., 2005), which are themselves both key in DNA double strand break repair. Complete loss of function of the protein product of either BRCA gene leaves cells extremely sensitive to PARP inhibitors presenting this therapeutic opportunity (Bryant et al., 2005; Fong et al., 2009) .

Other studies have highlighted a range of SSL interactions that may provide suitable targets for therapy (Bitler et al., 2015; Karnitz and Zou, 2015; Williamson et al., 2016). For example, PI5P4K kinases are essential in the absence of p53 (Emerling et al., 2014), inhibition of ENO2 inhibits viability in ENO1 deficient glioblastoma cells (Muller et al., 2012) and APE1 inhibitors in PTEN deficient cells results in the induction of apoptosis (Abbotts et al., 2014).

Currently, mainly due to experimental limitations (You et al., 2010) exhaustive experimental identification of human SSL interactions is not tenable. While here are many studies focused on screening for genetic interactions in model organisms (Stark, 2006) unfortunately, genetic interactions are not highly conserved between lower eukaryotes and their human orthologue equivalents (Wu et al., 2013). Instead, in order to identify novel human SSL interactions, we are left to infer and predict these pairs indirectly from existing human and model organism data through the use of models and other computational techniques (Benstead-Hume et al., 2017a).

Several classifiers have been developed to predict genetic interactions within model organisms. Wong et al. (Wong et al., 2004b) predicted genetic interactions in *S. cerevisiae* using decision tree classifiers with multiple data types and network topology. Paladugu et al. (Paladugu et al., 2008) focused on S. cerevisiae data; by extracting multiple features from protein interaction networks they achieved sensitivity and specificity exceeding 85% using support vector machine (SVM) classifiers. Later, Chipman et al. (Chipman and Singh, 2009) employed random walks and decision tree classifiers on protein interaction and gene ontology (GO) data to classify both S. cerevisiae and *C. elegans* negative genetic interactions.

Several classifiers have been developed to predict genetic interactions between species. Zhong and Sternberg (Zhong and Sternberg, 2006) classified *C. elegans* negative genetic interactions based on orthologous gene pairs in *S. cerevisiae* and Drosophila melanogaster. Jacunski et al. (Jacunski et al., 2015) developed SINaTRA (Species-INdependent TRAnslation) to classify *S. cerevisiae* SSL pairs based on *S. pombe* training data and vice versa, using features extracted from physical interaction data. The model trained on *S. cerevisiae* data was applied to predict 1,309 human SSL pairs with a reported false positive rate of 0.36. Similarly Wu et al. (Wu et al., 2014) developed MetaSL, an ensemble machine learning mode which applied eight different classifiers on *S. cerevisiae* data and applied it to predict human SSL pairs.

Using an alternative approach, the DAISY workflow predicted human SSL interactions directly from human cancer and cell–line data (Jerby-Arnon et al., 2014). The authors used somatic copy number variation and mutation profiles to achieve a ROC AUC score of 0.779 demonstrating a strong propensity (p-value < $1*10^{-4}$) for predicting SSL pairs in *H. sapiens*.

There are a number of additional recent studies that use biological networks to predict genetic interactions. Mashup (Cho et al., 2016) reported an average area under the precision curve (AUPR) of 0.59 for SSL and 0.51 for SDL pair prediction in a real human dataset. Others have utilised gene ontology terms to predict SSLs. These include Ontotype (Yu et al., 2016), where the authors predict the growth outcome on double knock-out of gene pairs. Their prediction set of gene pairs related to DNA repair and nuclear lumen correlated with Costanzo et al's (Costanzo et al., 2016) validated SSL dataset with a coefficient of r = 0.61. The authors of Dcell (Ma et al., 2018) constructed a visible neural network embedded in the hierarchical structure of 2526 subsystems describing the eukaryotic cell and used this to predict negative genetic interactions in *S. cerevisiae*.

In this study we introduce SLant (Synthetic Lethal analysis via Network topology), a random forest classifier trained on features extracted from the protein-protein interaction (PPI) networks of five species. These features comprise both node-wise distance and pairwise topological PPI network parameters and gene ontology data. Using SLant we provide in-species, cross-species and consensus classification for synthetic lethal pairs in all five organisms including human. We subsequently experimentally validated three of the predicted human SSLs in a human cell-line. Finally we identify a large cohort of candidate human synthetic lethal pairs which are available with the consensus predictions for all the model organisms in the Slorth database (http://slorth.biochem.sussex.ac.uk).

3.2 Materials and methods

3.2.1 Data Acquisition and pre-processing

Gene and orthology data were downloaded from Ensembl (Hubbard et al., 2002), Genetic interaction data were obtained from BioGRID (version 3.4.156) (Stark, 2006) with supplementary *D. melanogaster* data downloaded from Flybase (version 6.13) (Gelbart et al., 1996). Each gene was labelled with gene ontology (GO) data from the gene ontology consortium (Ashburner et al., 2000). Protein-protein interaction (PPI) data were obtained from the STRING database (version 10) (von Mering et al., 2005). To ensure reliability only experimentally derived and curated pathway data with a reliability cut-off of 80 were utilised (Supplementary Table 3.6). The Ensembl ENSP protein IDs in the PPI data sets were converted to their respective Ensembl ENSG gene IDs. This enabled us to relate the physical interaction data to the genetic interaction data and label each physical interaction gene pair as SSL (if present in the BioGRID data) or non-SSL (if the pair was not present in the BioGRID data).

For each organism an equal number of non-SSL pairs were assigned randomly to constitute the negative training set. When assigning a non-SSL pair, we checked to makes sure that its orthologues had not been assigned as having an SSL as, although it is not prescriptive, there is an enrichment of SSL pairs in orthologous genes.

Similar methods were used to build the training set used for our SDL interaction classifiers but we instead extracted BioGRID pairs annotated as synthetic dosage lethal as our positive class data.

3.2.2 Feature processing

The R (version 3.4.0) igraph package (version 1.1.2) (Csárdi and Nepusz, 2006) was used to generate a network representation of the PPI data for each of our 5 organisms and to calculate network features (Table 3.1). Whilst we extracted network features for just a subset of all possible gene pairs the entire network of protein interactions was used in each calculation.

Name	**Class**	**Description**
Betweenness	Node-wise	The number of shortest paths in the entire graph that pass through the node.
Constraint	Node-wise	Related to ego networks. A measure of how much a node’s connections are focused on single cluster of neighbours.
Closeness	Node-wise	The number of steps required to reach all other nodes from a given node.
Coreness	Node-wise	Whether a node is part of the k-core of the full graph, the k-core being a maximal sub-graph in which each node has at least degree k.
Degree	Node-wise	The number of edges coming in to or out of the node.
Eccentricity	Node-wise	The shortest path distance from the node farthest from the given node.
Eigen centrality	Node-wise	A measure of how well connected a given node is to other well-connected nodes.
Hub score	Node-wise	Related to the concepts of hubs and authorities the hub score is a measure of how many well linked hubs the nodes is linked to.
Neighbourhood n size	Node-wise	The number of nodes within n steps of a given node for n of 1, 2, 5 and 6
Adhesion	Pairwise	The minimum number of edges that would have to be severed to result in two separate sub-graphs separating the source and target nodes.
Cohesion	Pairwise	The minimum number of nodes that would have to be removed to result in two separate sub-graphs separating the source and target nodes.

Adjacent	Pairwise	Whether a source and target node are connected via an edge.
Mutual neighbours	Pairwise	How many first neighbours a target and source node share.
Shortest path	Pairwise	The minimal number of connected vertices that create a path between the source and target node.
Between community	Pairwise	A logical feature stating whether the source and target nodes inhabit the same community produced by the spin glass random walk.
Cross community	Pairwise	A logical feature stating whether the source and target nodes connect two communities as produced by the spin glass random walk.
Shared GO count – Biological process	Go term	The number of biological process GO annotations shared between the source and target node.
Shared GO count – Molecular function	Go term	The number of molecular function GO annotations shared between the source and target node.
Shared GO count – Cellular compartment	Go term	The number of cellular compartment GO annotations shared between the source and target node.

Table 3.1. Names and descriptions of the node-wise and pairwise network parameters and GO term features used in Slant.

The features generated for our models were broadly categorised as node-wise or pairwise features as listed in Table 3.1. In general node-wise features, such as degree, were calculated by extracting network parameters for single nodes and finding the averaged distance between them as a pairwise feature. Pairwise features such as shortest path were calculated by igraph on each pair. To calculate shared GO terms, classed as a pairwise feature, we took a count of overlapping GO terms between the genes in each pair.

To generate our community features we applied a spin-glass random walk using the R igraph communities module to assign genes to 20 distinct communities separated by choke points across the graph. The final count of communities, 20, was chosen by measuring the predictive performance of our community features with a community count incrementing in steps of 5. After 20 communities we saw no further improvement.

The entire feature generation pipeline for the full complement of available gene pairs proved computationally intense, especially the generation of pairwise features such as cohesion, and run-time took up to 120 hours for each organism on an 8x Intel Xeon 3.50GHz processor with 16Gb RAM.

3.2.3 Training and test sets

Before analysis all features in each dataset were normalised so that all feature values fell between 0 and 1. The resulting feature sets were divided into training, test and unlabelled sets. For each organism the feature set was under sampled to provide a balanced training set with an equal number of SSL and non-SSL pairs. The training set was further partitioned 80:20 to create a test set. The non-SSL pairs removed from the training data as part of under sampling were set aside as unlabelled data to be used in the prediction section of this study .

3.2.4 Creating balanced training and test pair sets with distinct gene components

Some genes are highly represented in our available SSL training data whilst some only occur once, so generating two sets with balanced classes and a requisite number of observations posed a challenge. To create balanced training and test datasets with enough observations to perform validation we first created a list of genes ranked by the number of pairs they were found in. Next we divided this list adding the first to our list of genes available in our training data, the second to our test data and so on so that both data sets had a similar distribution of gene representation. Finally we used these two gene lists to filter our feature data into two subsets with no overlapping genes and balanced class.

3.2.5 Analysis and modelling

We used the “ranger” e1071 random forest classifier, part of the R caret library, to model and classify SSL and non-SSL interactions in our training set. 5-fold cross validation was applied to each organism's training set to tune the model's hyper-parameters and the best model was used to assess predictive performance within each species. These optimised models were then used to predict SSL pairs across species, both in *H. sapiens* and across all other model organisms. These predictions were outputted as the probability of each class and were validated against the test data set.

3.2.6 Calculating cross species consensus

In an attempt to further improve accuracy, as well as pairwise cross-species predictions, a consensus was taken from the predictions on the test set from all other species. This consensus was calculated by running a second classifier, a boosted Generalized Linear Model (GLM) that was trained on the previous classifiers outputs. To allow for validation this consensus dataset was segmented into a train and test set (both 0.5 the size of the original due to the smaller overall size). Finally we used this consensus model to predict SSL pairs in the unlabelled data set.

All of the R source code for Slant is available publicly at https://bitbucket.org/mrgraeme/slant. All data used is available via public sources.

3.2.7 Validation using clonogenic survival assays

A subset of potential SSL interacting pairs featuring PBRM1 (BAF180) complemented with genes with a known inhibitor were chosen from our predictions for experimental validation (Supplementary Table 3.5).

3.2.7.1 Cell culture

U2OS-derived control and PBRM1-deficient cell lines (Hopkins and Groom, 2002; Hopkins et al., 2016) were cultured in Dulbecco DMEM supplemented with 10% FBS, glutamine and Penicilin/Streptomycin.

3.2.7.2 Clonogenic survival assays

Cells were seeded and allowed to adhere prior to drug treatment. Cells were exposed to the indicated amount of drug in triplicate, and incubated for 14 days at 37C with 5% CO2 prior to staining with methylene blue ((0.4%). Cell colonies were manually counted and presented as the surviving fraction relative to the untreated cells.

3.3 Results

A genome-wide protein-protein interaction (PPI) network was constructed for H. sapiens and each of our model organisms (S. cerevisiae, *D. melanogaster*, *C. elegans*, and S. pombe) using PPI data from the STRING database (von Mering et al., 2005). In this network, each node represents a protein and each edge represents a physical interaction between two proteins. For each pair of proteins 12 node-wise and 7 pairwise features were extracted from the PPI network using the R igraph library (Csárdi and Nepusz, 2006). Each protein in the network was labelled with its respective Ensembl gene identifier so that this physical interaction data could be matched with gene interaction data. For each gene pair 3 additional GO term related features were generated using Gene ontology (GO) data (Ashburner et al., 2000).

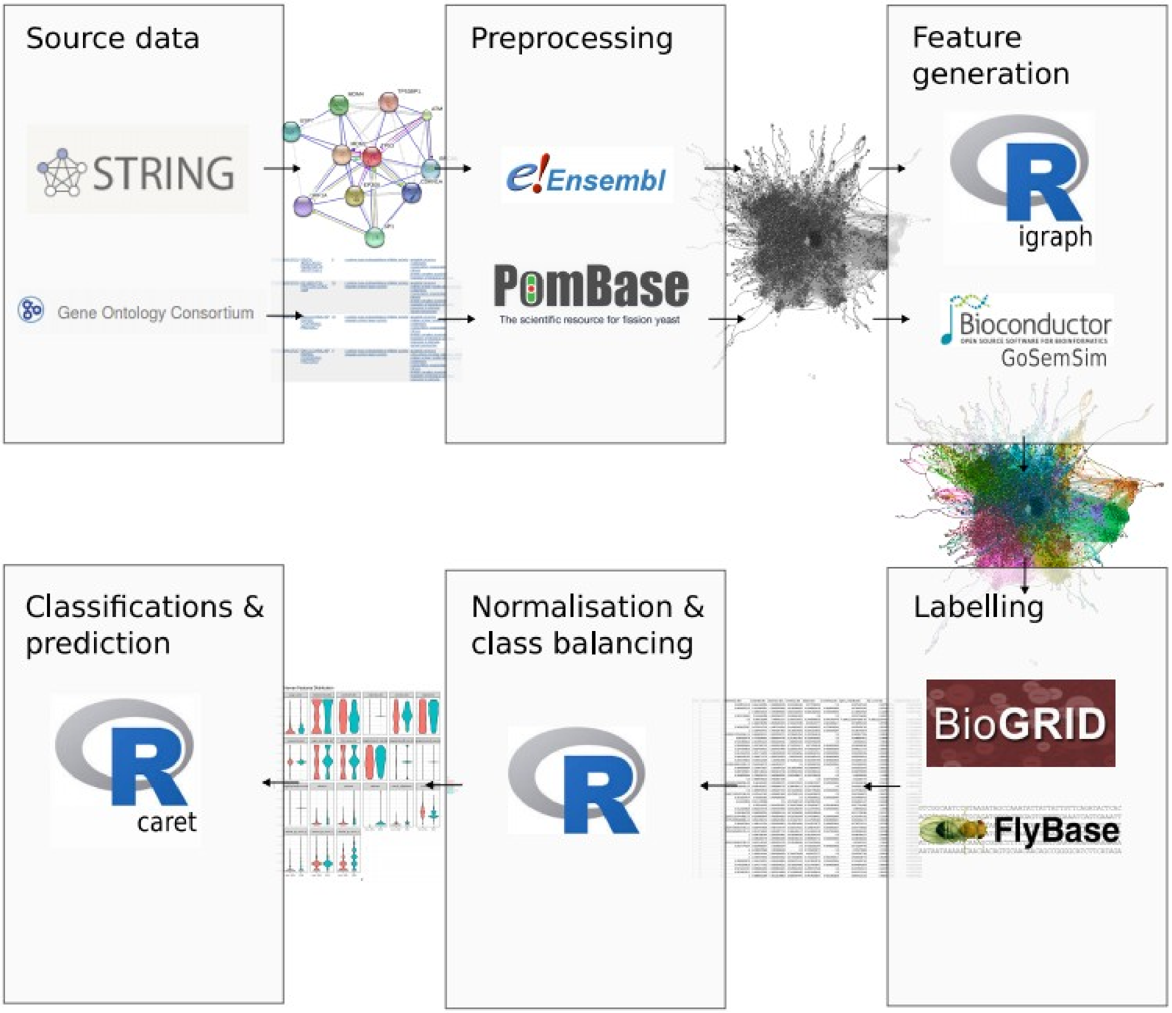


Figure 3.1. A schematic visualising how SLant's source data is collated from STRING and the Gene Ontology Consortium, preprocessed so that this source data can be directed joined with BioGRID data for labeling and processed to create the final training set. Feature generation was completed using R, the R igraph library and GoSemSim, a Bioconductor package.

For each PPI network, pairs of proteins whose respective genes were identified as having a negative genetic interaction in BioGRID (Stark, 2006) were labelled as having an SSL interaction (Figure 3.1). Equal numbers of SSL and non-SSL gene pairs were selected independently for the training sets for each species (see methods). Similarly we created training sets for SDL and non-SDL gene pairs in *H. sapiens* and *S. cerevisiae*, the only two species where there is enough data for prediction purposes.

3.3.1 Network parameter distributions in humans

The features used for classification in the SLant algorithm were broadly divided into node-wise, pairwise or GO-term related categories. Node-wise features were derived from an individual node's network parameter, such as degree or centrality. These node-wise features were converted to pairwise features by taking the average distance for that feature between the nodes in each pair. Pairwise features were defined as those that apply to a pair of nodes such as shortest path or cohesion. The spin glass random walk features discussed below were included in our pairwise category. GO related features were derived from shared annotations between pairs of genes (Ashburner et al., 2000) (for a full list of features see Table 3.1).

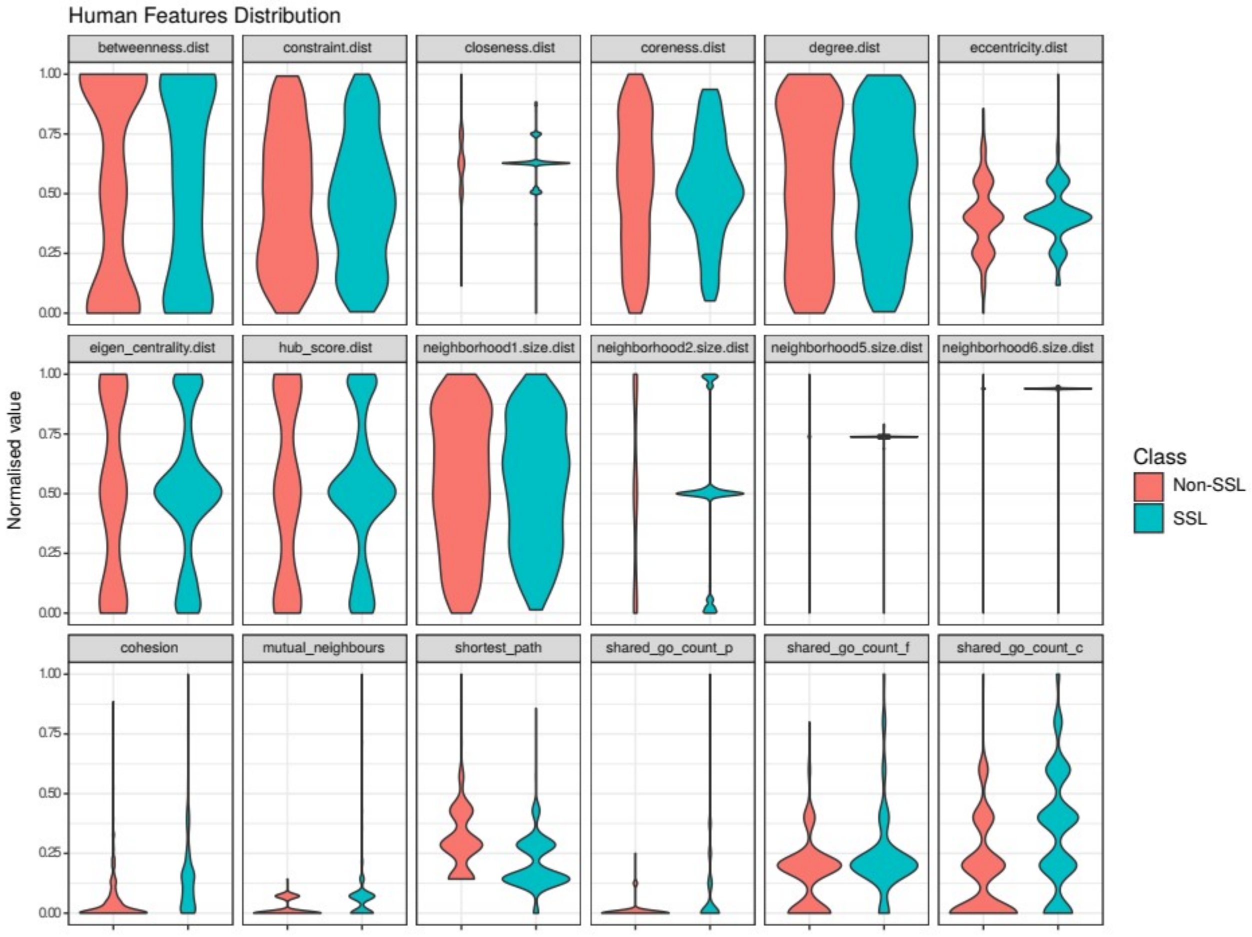


Figure 3.2. A set of violin plots illustrating the value distributions for each feature in our human training set grouped into SSL and non-SSL classes. The features were derived from 411 SSL and 411 non-SSL gene pairs (see Supplementary Table 3.6). Feature distributions that show greater variance between SSL and non-SSL gene pair classes, for example the shortest path feature, often provide improved predictive power in classifiers.

Figure 3.2 shows the distribution of these features in SSL and non-SSL gene pairs in humans. In general pairwise parameters showed a greater variance between SSL and non-SSL classes than our node-wise parameters suggesting they may prove better predictors in our models. Of these pairwise parameters the most notable differences were observed in the parameters labelled: cohesion - the minimum number of nodes that would have to be removed to result in two separate sub-graphs separating the source and target nodes, shortest path - the minimum number of nodes that must be traversed in a path between the source and target gene, and mutual neighbours - the number of nodes that are shared as neighbours between the source and target gene.

The higher values exhibited by gene pairs in the SSL class for the cohesion feature (paired t-test; p = $2.2*10^{-16}$ in *H. sapiens*) suggest that SSL pairs are generally more densely connected in a physical interaction graph than non-SSL pairs (Supplementary Figure 3.1a).

We also note that the shortest path between gene pairs is shorter on average for SSL gene pairs compared to non-SSL gene pairs (paired t-test; p = $4.589*10^{-11}$ in *H. sapiens*) (Supplementary Figure 3.1b) and, related to the shortest path parameter, SSL genes more often share a large number of mutual neighbours (paired t-test; p = $4.058*10^{-11}$ in *H. sapiens*) (Supplementary Figure 3.1c).

In terms of node-wise features it is of some interest to note that the difference between neighbourhood sizes of two genes in an SSL pair often differ more than those in a non-SSL pair.

3.3.2 Random walk community generation suggests that most SSL interactions occur between rather than within clusters of genes

In an attempt to ascertain whether synthetic lethal interactions occurred within or between local clusters of genes in our physical interaction network we applied a spin-glass random walk to assign genes to 20 distinct communities separated by choke points across the graph (Figure 3.3a). Analysis showed that the majority of SSL interactions occurred between these communities rather than within (Figure 3.3b). In addition pairwise topological analysis suggests that SSL pairs of genes have shorter paths between them than non-SSL pairs and a higher occurrence of adjacency. Together these analyses suggest that SSL pairs are often at the peripheries of these communities, connecting their respective clusters.

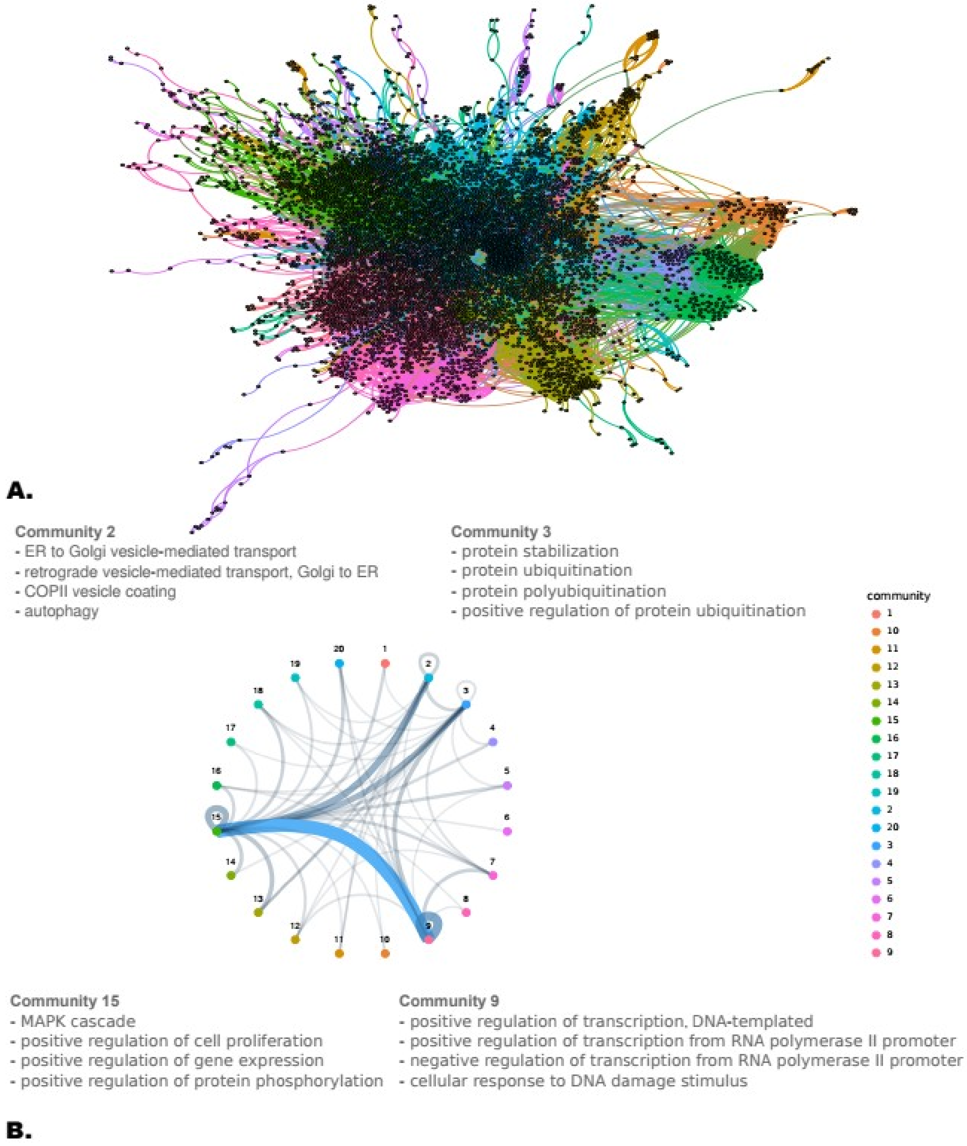

A.
Community 2
- ER to Golgi vesicle-mediated transport
- retrograde vesicle-mediated transport, Golgi to ER
- COPII vesicle coating
- autophagy
Community 3
- protein stabilization
- protein ubiquitination
- protein polyubiquitination
- positive regulation of protein ubiquitination
community
1
10
11
12
13
14
15
16
17
18
19
2
20
3
4
5
6
7
8
9
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Community 15
- MAPK cascade
- positive regulation of cell proliferation
- positive regulation of gene expression
- positive regulation of protein phosphorylation
Community 9
- positive regulation of transcription, DNA-templated
- positive regulation of transcription from RNA polymerase II promoter
- negative regulation of transcription from RNA polymerase II promoter
- cellular response to DNA damage stimulus
B.

Figure 3.3. a. Human protein-protein interaction network with clustered communities generated by a spin glass random walk. Nodes and edges are coloured by their source community cluster as per the legend provided in Figure 3.3 b. Figure 3.3 b. Community cluster connection graph where the weight of each connection corresponds to how many SSL interacting pairs begin and end at each community. We observe the largest count of SSL interactions occurring between cluster 9, *notably associated with transcription regulation and DNA damage response GO terms and cluster 15, associated with MAPK cascade, cell proliferation and gene expression GO terms.*

Based on these observations we were able to create two additional features which provide further predictive power for classifying SSL pairs; whether nodes shared a community and whether the pair connected two communities.

3.3.3 SSL pairs shared more GO annotations than non-SSL pairs

The count of shared GO terms, that is the number of GO annotations that two genes in a pair share with each other, also varies between SSL and non-SSL observations. SSL pairs generally share, on average, less biological process GO annotations (Supplementary Table 3.1) than non-SSL pairs ($p < 2.2*10^{-16}$ in *H. sapiens*) and proportionately more molecular function and cellular component GO annotations ($p < 2.2*10^{-16}$ in *H. sapiens* for both biological process and cellular compartment terms). This supports the view that that SSL protein product pairs are often found in similar but distinct pathways rather than within a single pathway (Kelley and Ideker, 2005). Damaging two complementary functional pathways is likely cause more stress to the cell than damaging one pathway twice and leaving the complementary pathway functional.

Although the GO annotation based features above provide predictive power in our models as discussed below, due to the hierarchical nature of GO annotation, comparing the absolute count of shared GO terms does present some issues. As such GoSemSim (Yu et al., 2010) was used to further measure the semantic similarity between SSL and non-SSL pairs. We found that in *H. sapiens* SSL pairs showed a significantly higher semantic similarity score (mean = 0.65) that non-SSL pairs (mean = 0.57) (Welch two sample t-test $p=4.6*10^{-7}$).

Analysis of GO terms present in paired SSL genes found that the most commonly shared molecular functional GO annotation related to protein binding (Supplementary Figure 3.2). Other molecular function GO annotations commonly found associated between SSL pairs include protein complex binding, GTP binding, DNA binding and GTPase activity. At the level of biological process GO annotation for SSL gene pairs we also noted associations with terms related to positive regulation of cell proliferation and negative regulation of apoptotic process as well as those labelled with positive regulation of gene expression and positive regulation of transcription from RNA polymerase II promoter.

In an attempt to further quantify the GO annotation driving the variation between genes found in SSL pairs and those not found in SSL pairs we employed a GO enrichment analysis using the on-line GOrrila tool (Eden et al., 2009). We found significant enrichment in a number of GO annotations including negative regulation of cell differentiation ($p = 9.15*10^{-3}$), positive regulation of transcription by RNA polymerase II ($p = 9.53*10^{-3}$) and regulation of Notch signalling pathway ($p = 8.85*10^{-3}$) in the biological process ontology but no further enrichment in the molecular function or cellular compartments ontologies. All p-values have been are corrected for false positives using the Benjamini Hochberg method

3.3.4 SSL interactions in essential genes

Comprehensive studies of *S. cerevisiae* genetic interactions by Costanzo et al. (Costanzo et al., 2010, 2016) have found that essential genes that share an edge on the PPI network are enriched for genetic interactions and that is consistent with previous observations (Kelley and Ideker, 2005). As our classifiers in part use the distance of gene pairs as a predictive feature we performed analysis to ensure our predictions were not simply picking out gene pairs enriched for essential genes.

We first noted that the range of shortest path values between SSL pairs on the protein-protein interaction (PPI) network runs from 1 to 7 with a mean of 2.43 and a standard deviation of 0.78 affirming that our training set features many SSL pairs that are not adjacent in the PPI network.

Using a set of essential human genes defined by Wang et al. (Wang et al., 2015), we found that 11% of the genes in our SSL training set were defined as essential, where as for non-SSL genes it only 0.7%. For human gene pairs ~1.7% of SSL pairs and ~1.4% of non-SSL pairs are comprised of two essential genes. We also found that 29% of SSL pairs and 22% of non-SSL pairs included at least one essential gene.

Upon comparison we found that ~22.5% of our SSL predictions included at least one essential gene and ~1.4% featured two essential genes, a ratio comparable with our training data. This suggests that our predictions are not further enriched for essential genes.

3.3.5 Models explaining patterns of genetic interactions

There are three models used to explain how genetic interactions occur (Hin et al., 2004; Kelley and Ideker, 2005; Tong et al., 2001b). The “between pathway model” is where the genetic interaction involves genes in two distinct pathways with complementary functions. A deletion of a gene in one pathway abrogates the function of that pathway and the cell cannot survive with of both pathways are lost. The “within a pathway model” is where genetic interaction occurs between genes in the subunits of a single pathway. Loss of one gene can be tolerated but the additive effects of the loss of several genes in that pathway are lethal. Finally ‘the indirect model’ is where the phenotype is not mediated by a localised mechanism.

Previous computational analyses have found that negative genetic interactions are enriched both between biological processes (or pathways) and within biological processes, giving credence to these models (Costanzo et al., 2010, 2016; Kelley and Ideker, 2005; Ulitsky and Shamir, 2007). SSL interactions occur primarily between local clusters in the PPI network suggest that the between pathways interactions may still involve pathways that are close in PPI space. This may explain why the analysis of PPIs is so effective in predicting SSL interactions.

3.3.6 Network parameter distributions in model organisms

The distribution of network parameters across our four model organisms widely followed similar trends with our human feature set. Again the pairwise features for each organism appear to vary more between SSL and non-SSL classes than node-wise features. A few dissimilarities were noticeable, for example while SSL gene pairs tend to exhibit a higher levels of adhesion and cohesion in H. sapiens, S. cerevisiae (Supplementary Figure 3.3a) and *D. melanogaster* (Supplementary Figure 3.3b) the distribution for these features were notably inverted in *C. elegans* (Supplementary Figure 3.3c) and S. pombe (Supplementary Figure 3.3d) so that non-SSL pairs showed higher adhesion and cohesion than SSL pairs.

3.3.7 Validating SSL gene pair classification

In this study we perform two classifications. First in-species classification, classifying and validating SSL gene pairs using training and test data from the same organism. Then cross-species classification where we use the models built using the training data for each organism to blindly predict SSL for each other species. Within each species, the feature data were normalised and segmented into training and test sets with 20% set aside for validation.

		Validation results				
		H. sapiens	S. cerevisiae	C. elegans	D. melanogaster	S. pombe
Model	H. sapiens	0.965	0.722	0.415	0.736	0.64
	S. cerevisiae	0.736	0.835	0.754	0.728	0.58
	C. elegans	0.525	0.716	0.682	0.759	0.47
	D. melanogaster	0.725	0.744	0.744	0.873	0.72
	S. pombe	0.657	0.69	0.637	0.756	0.83
	Consensus	0.985	0.853	0.360	0.883	0.66

Table 3.2. *Cross validation ROC AUC scores for each organism from both in-species and cross species SSL models. The best score for each species model is highlighted in green. Models are displayed vertically in rows with the consensus model displayed at the bottom of the table and the results for those models are displayed in columns with the consensus results highlighted in blue.*

We employed 5-fold cross validation to optimise the hyper-parameters for each organism's random forest classifier and evaluated in-species classification performance (Table 3.2). In this study our random forest classifiers utilised just one hyper-parameter, mtry - the number of variables randomly sampled as candidates at each split for each tree. The best classifier for each species was then used to predict the SSL gene pairs in each of the other four species. Table 3.2 shows the ROC AUC scores for both the in-species and cross-species predictions for all of our models.

Although it is difficult to compare the performance of classifiers due to varied validation sets, the ROC AUC score of 0.965 for *H. sapiens* SSL gene pair classification achieved by the Slant classifier (using holdout validation data) appears to out-perform Daisy's ROC AUC score of 0.779.

Our initial in-species classification of S. cerevisiae SSL resulted in relatively low performance (AUC 0.734) compared to other related studies. For example MetaSL, which used a much smaller data set of just 7,347 SSL pairs compared to Slant's 395,199 pairs, achieved ROC AUC scores of up to 0.871 (Wu et al., 2014). In order to mitigate any noise or error introduced in our large dataset we filtered out any SSL interactions reported in BioGRID supported by less than 3 supporting publications for S. cerevisiae and less than 2 papers for S. pombe. Our training data ultimately used 17,568 out of a total 395,199 SSL pairs available for S. cerevisiae and 3,836 out of 35,391 SSL pairs for S. pombe. These sample sizes should still be large enough to generalise well for out of sample predictions as well as preforming well in classification and validation. Filtering our yeast data improved our scores from AUC ROC 0.734 to AUC ROC 0.883 for S. cerevisiae and 0.728 to 0.889 for S. pombe which suggests that by removing pairs that show fewer citations in the BioGRID data we are reducing variation in our training data introduced by false positives. This may be due to the relatively high false-positive rate found in large scale GI screenings, an observation supported by analysis performed by Campbell & Ryan et al. Who estimated that large scale screenings can suffer a false positive rate of up to ~10% (Campbell et al., 2016). Using this value we can calculate that by removing GI pairs with less than 2 and 3 references respectively we may be reducing false positive rates from 1/10 to 1/100 in S. pombe and from 1/10 to 1/1000 in S. cerevisiae.

Cross-species predictions of SSLs were quite variable in performance. Models from both *S. cerevisiae* and *D. melanogaster* and *C. elegans* were successful in predicting human SSLs with AUC ROC scores of 0.713, 0.727 and 0.769 respectively.

Although the *C. elegans* classifier performed relatively poorly in our cross-species validation for H. sapiens classification, this variation may help improve the generalisation of our consensus model which is discussed below. To test this cross-species validation was performed without the worm model. The removal of worm data from the classifier resulted in a small but noticeable decrease in performance of the consensus classifier for humans (decreasing from ~0.985 to ~0.92).

The result here suggest that the PPI patterns between SSL genes are similar both within and between species and that network topology features used in our classifiers generalise well across organisms. We identified the most predictive features for each organism and found that the same features were most predictive in many of the species. The shared GO count features were important in all organisms except S. pombe and the pairwise features adhesion, cohesion, mutual neighbours and adjacency were important in all organisms except *C. elegans*. Two node-wise features, coreness and neighbourhood size are also listed as important features across 3 organisms (Supplementary Table 2).

3.3.8 Class balance changes do not significantly impact classifier performance

As described below in methods each of these models use a balanced training set with a ratio of 1:1 interacting and non-interacting pairs, however in reality the ratio between interacting and non-interacting pairs is likely more in the order of 1:50 based on global yeast GI screens (Costanzo et al., 2016). To ascertain that our class balance has not unduly biased our prediction in any way we re-ran our classifiers using a randomised training / validation set with approximately 1:10 and 1:50 class balance. We found that with a class balance of 1:10 our performance remained stable and with a class balance of 1:50 we found just a small drop in performance (human AUC ROC ~0.87 compared to the original ~0.965 and consensus AUC ROC ~0.90 compared to ~0.985).

3.3.9 Our models are robust to incompleteness in the source PPI networks

It is known that our current PPI models are incomplete (Huttlin et al., 2017; Mosca et al., 2013) and suffer from ascertainment bias. That is, some genes, and indeed some species, are better studied than others. To test our model's robustness to the incomplete nature of the protein-protein interaction networks, we re-ran our classifiers holding out 10% and 20% of the nodes, at random, from original PPI data in *H. sapiens*. In the case of the 90% 'complete' PPI network the performance of our in-species model validation was not effected and our *H. sapiens* consensus showed just a small drop in performance (from AUC ROC ~0.985 to ~0.922). With a 80% 'complete' *H. sapiens* PPI network we saw another fairly small incremental drop in *H. sapiens* consensus performance (AUC ROC ~ 0.85) and a small drop in *H. sapiens* in-species performance (AUC ROC dropping from 0.965 to 0.911). This suggests both that an increasingly complete PPI network may incrementally improve our predictive performance and that the current models are fairly resilient to the incomplete nature of the PPI network.

3.3.10 Our pair-wise distance features are the most predictive

In addition to the feature importance analysis performed in this study we also re-ran our classifiers holding out our 12 node-wise distance features, 6 pair-wise features and 3 GO-term related features in turn. We found that the model holding out pair-wise features saw the largest drop in performance in consensus with the *H. sapiens* consensus ROC AUC dropping from ~0.985 to ~0.730 and the in-species *H. sapiens* ROC AUC dropping from ~0.965 to ~0.82. In comparison to our models holding out node-wise features saw a more notable drop in the in-species performance (*H. sapiens* consensus ROC AUC dropping from ~0.985 to ~0.85 and in-species *H.sapiens* from ~0.965 to ~0.823). Similarly holding out our GO term features resulted in a decrease in predictive performance (*H.sapiens* consensus ROC AUC dropping from ~0.985 to ~0.882 and in-species *H.sapiens* from ~0.965 to ~0.890).

3.3.11 Our models are moderately robust to pair-input bias

As discussed by Park et al. (Park and Marcotte, 2012) computational prediction methods that utilise gene pair observations, such as the models in this study, can be subject to positive bias in validation. They discovered that model validation performed significantly better when genes that made up the pairs in the test set were also featured in the training set compared to those models where they were not.

In order to evaluate how Slant's validation was effected by pair-input bias we generated a test set from our raw feature data in which none of the genes featured in the test pairs were present in any of the pairs featured in the training set. We refer to these as segregated datasets.

To make sure we could make a fair comparison we generated a further control training and test set by randomly sampling the pairs created above from both segregated data sets. This ensured that the pair count and the pairs themselves remained the same while gene components could be shared between our control training and test sets.

Running our models again using these segregated training and test data we achieved a AUC ROC of 0.789 for predicting human SSL pairs, compared to 0.845 for our control datasets and 0.965 for our full training and test sets. This suggests that while our predictions may be somewhat biased towards genes that are featured in the training data our models also appear to predict SSL pairs comprised of genes that are not in our training data and, more importantly, potentially genes that have not previously been associated with SSL interactions.

3.3.12 A consensus based on many cross-species predictions further improves performance

To further expand our model we took a consensus from the cross-species predictions for each organism. This consensus was calculated by running a second classifier, a boosted general linear model (GLM) that was trained on the previous cross-species classifier output. This output took the form of confidence scores. For example, for any particular pair of human genes the confidence scores given to that pair by every cross-species classifier were used as features. The probability outputted by this final classifier is referred to as the consensus score.

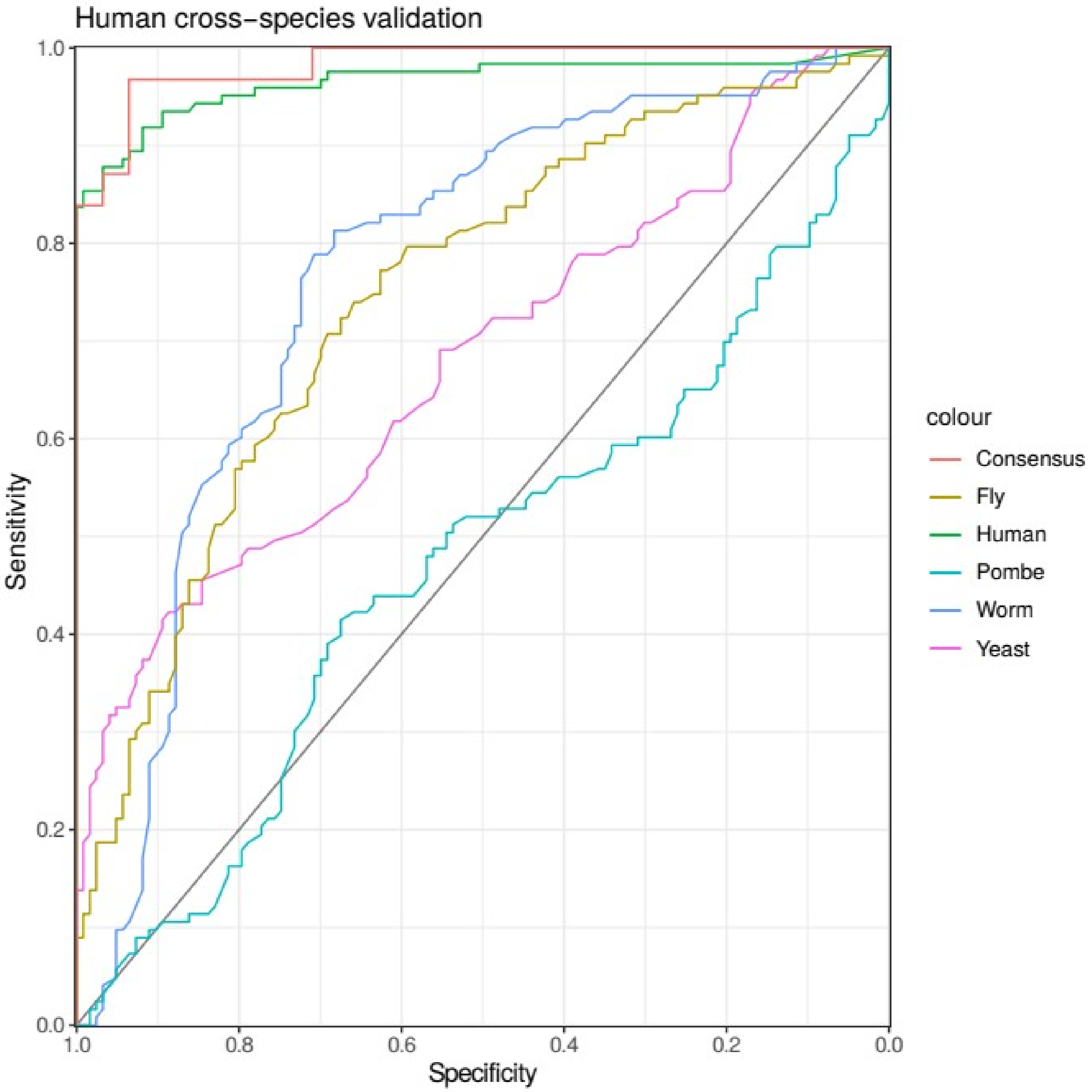


Figure 3.4. Cross-species ROC AUC scores for each models classification performance on our human SSL interaction validation set. An additional curve for our consensus predictions was added separately based on the performance of the consensus validation set.

To allow for validation this consensus dataset was segmented into a training and test set (both 0.5 the size of the original due to the smaller overall size). The ROC AUC for our consensus prediction validation was also plotted and achieved a score of up to 0.985 when predicting *H. sapiens* SSL pairs, a further improvement on our in-species human validation ROC AUC score of 0.965 (Figure 3.4).

3.3.13 Predicting synthetic dosage lethal pairs

To ascertain whether SSL and synthetic dosage lethality (SDL) interactions share topological predictors we re-purposed our models to predict SDL gene pairs. We achieved an in-species AUC of 0.78 for *H. sapiens* pairs and 0.89 for *S. cerevisiae* pairs, a significantly improved score compared to that achieved during *S. cerevisiae* SSL pair classification. Our consensus model, utilising just *H. sapiens* and *S. cerevisiae* data, improved our *H. sapiens* predictions slightly (ROC AUC 0.80) (Supplementary Table 3.3).

SDL and SSL pairs in *H. sapiens* exhibit broadly similar feature distribution and feature importance for both classifiers. Despite this only 7,531 pairs were predicted as both SDL and SSL (of 41,103 SDL pair predictions and 59,475 SSL pair predictions).

In our human SDL models cohesion and shared cellular compartment GO terms featured as important features for both classifiers though molecular functional GO term annotation proved an important feature for SDL classification while shared biological process GO term featured well for SSL classification. The closeness feature, which measures how many steps is required to reach all other nodes from a given node, performed well for SDL classification. On the other hand coreness, a measurement of how well connected a node's neighbours are compared to the graph overall provided better predictive power for SSL classification.

We next compared biological process GO terms present in SDL and SSL pairs. We found that DNA damage related processes were more frequently seen in SDL pair data than in SSL pair data (~1.00% cellular response to DNA damage stimulus, ~0.70% DNA repair in SDL pairs compared to ~0.53% and ~0.46% respectively in SSL pairs). MAPK cascade and regulation of cell proliferation processes were well represented in both groups.

3.3.13.1 Comparison to previous studies

As discussed in the introduction, a number of other studies have used similar methods to predict genetic interactions. Most notably, this study shares a number of similarities with SINaTRA (Jacunski et al., 2015). However, Slant has been developed for a wider number of organisms, including using human data directly, uses an enhanced feature set, our predictions have been experimentally validated (see below) and all of our data are available via the Slorth database (see below).

Algorithmically, the similarities between Slant and SINaTRA include some of the features used and the treatment of normalisation to allow cross-species prediction. However the PPI data used by Slant were sourced from STRING and were filtered for reliability, while SINaTRA's PPI data were sourced from BioGRID. A number of key algorithmic differences include Slant's use of consensus models, for both SSL and SDL interactions, and the use of a large range of topological, community and GO features. Slant also treats node-wise features differently and includes the averaged difference between genes in a pair as well as the individual values for each gene. We show that the novel features present in Slant improve the results in the feature holdout section (see Our pair-wise distance features are the most predictive) and propose that the different data sets appear to be providing a large impact on the results. A comparison of the features used in the two studies are available in Supplementary Table 3.7.

Unfortunately, the source code for SINaTRA is not available. However we were able to assess how our algorithm performed compared to SINaTRA, by testing it on the historical yeast SSL data from BioGrid 3.2.104 that had been used in the development of the SINaTRA algorithm. SINaTRA reports impressive AUC ROC values of 0.92 for in-species *S. cerevisiae* SSL predictions, 0.93 for in-species *S. pombe* SSL predictions, 0.86 for *S. cerevisiae* to *S. pombe* cross species validation and 0.74 for *S. pombe* to *S. cerevisiae* cross species validation. We obtained similar results using cross validation (as reported by SINaTRA) with AUC ROC values of 0.98 for in-species *S. cerevisiae* SSL predictions, 0.98 for in-species *S. pombe* SSL predictions, 0.88 for *S. cerevisiae* to *S. pombe* cross species validation and 0.77 for *S. pombe* to *S. cerevisiae* cross species validation (see Supplementary Table 3.8).

Next, we re-implemented SINaTRA by running Slant with a close approximation of the features that SINaTRA used originally but using the current STRING PPI network and current SSL data for training (see Supplementary Tables 3.9 and 3.10). We found that Slant outperformed SINaTRA in all tests apart from the S. pombe to S. cerevisiae cross species validation (AUC ROC 0.607 versus 0.609). In particular Slant considerably outperforms SINaTRA using models generated using the pair-wise non-bias segregated training sets. This supports our theory that the additional pairwise features incorporated into Slant leads to a generalisation of the models.

Finally we analysed the 2518 predicted human SSL pairs, with a SINaTRA score of over 0.90, that were published in the original paper. Of these, none of these predictions have subsequently been reported in BioGRID, either as SSLs or as negative genetic interactions. However, the number of reported SSLs for humans is still rather low. Encouragingly, 55% of the SINaTRA high confidence SSL predictions were also predicted to be SSLs by Slant.

3.3.14 Slorth database

We employed the full cross-species consensus model to predict SSL and SDL gene pairs in all of our species. All pairs that did not achieved a consensus score of over 0.75 were removed from our final prediction list. All predictions are available in the Slorth database http://slorth.biochem.sussex.ac.uk.

The graphical visualizations of the SSL predictions and the experimentally derived SSL interactions from our training data (Supplementary Figure 3.4a) shows that the SSL network becomes much denser around the genes represented in the initial training data from BioGRID. This suggests that genes already implicated in an SSL pairs may share more SSL interactions than currently experimentally identified.

3.3.15 Predicting and validating SSL gene pairs associated with cancer

Using the models and classifiers described above we have identified and validated previously unpublished human SSLs that could be exploited therapeutically in the treatment of cancer. To identify potential therapeutic targets using our consensus method, we identified all the SSL gene pairs in *H. sapiens* where one of the genes had been identified as a tumour suppressor by the cancer gene census (Futreal et al., 2004) (Supplementary Figure 3.4b, Supplementary Table 3.4) and the other was a target of a drug approved for human use.

We found an enrichment in highly scoring SSL pairs containing the tumour suppressors VHL and PTEN. SSL pairs with the highest consensus scores included SREBF1, a transcription factor that binds to sterol regulatory element-1 and VHL (confidence score 0.810) and PTEN and SFN, a gene associated with breast cancer (confidence score 0.808). Other novel, highly scoring gene pair predictions that included cancer associated genes included PARP1 with PBRM1, BRCA2, ARID1A and APC as well as PIK3CA with MAP2K1, ABL1 and EGFR.

Validation on a handful of these predicted pairs providing some evidence that PBRM1 / PARP1 and PBRM1 / ABL1 share previously undescribed SSL interactions. We also see some evidence that PBRM1 / POLA1 share a synthetic rescue interaction.

3.3.16 Experimental validation of predictions

A set of predicted gene pairs, where one of the genes identified was PBRM1, was selected for experimental validation. The PBRM1 gene codes for the tumour suppressor BAF180 a protein that plays a key role in both chromatin remodelling and gene transcription. It is frequently mutated in a subset of cancers including Clear Cell

Papillary Renal Cell Carcinoma and Clear Cell Renal Cell Carcinoma (Brownlee et al., 2012) We chose gene pairs where the second gene codes for a protein which has published inhibitors. These included; PARP1, ERBB2, RAF1, POLA1, JAK2, ABL1, GSK3B (Supplementary Table 3.5). Inhibitors were chosen and procured via Sellekchem (https://pubchem.ncbi.nlm.nih.gov/source/Selleck%20Chemicals).

Clonogenic survival assays (Franken et al., 2006) were prepared for a control group and a BAF180 knockout group from the U2OS cell line. Both cell groups were treated with a range of drug concentrations based on previous literature for each. The resulting cell colonies were stained and counted after 14 days of incubation.

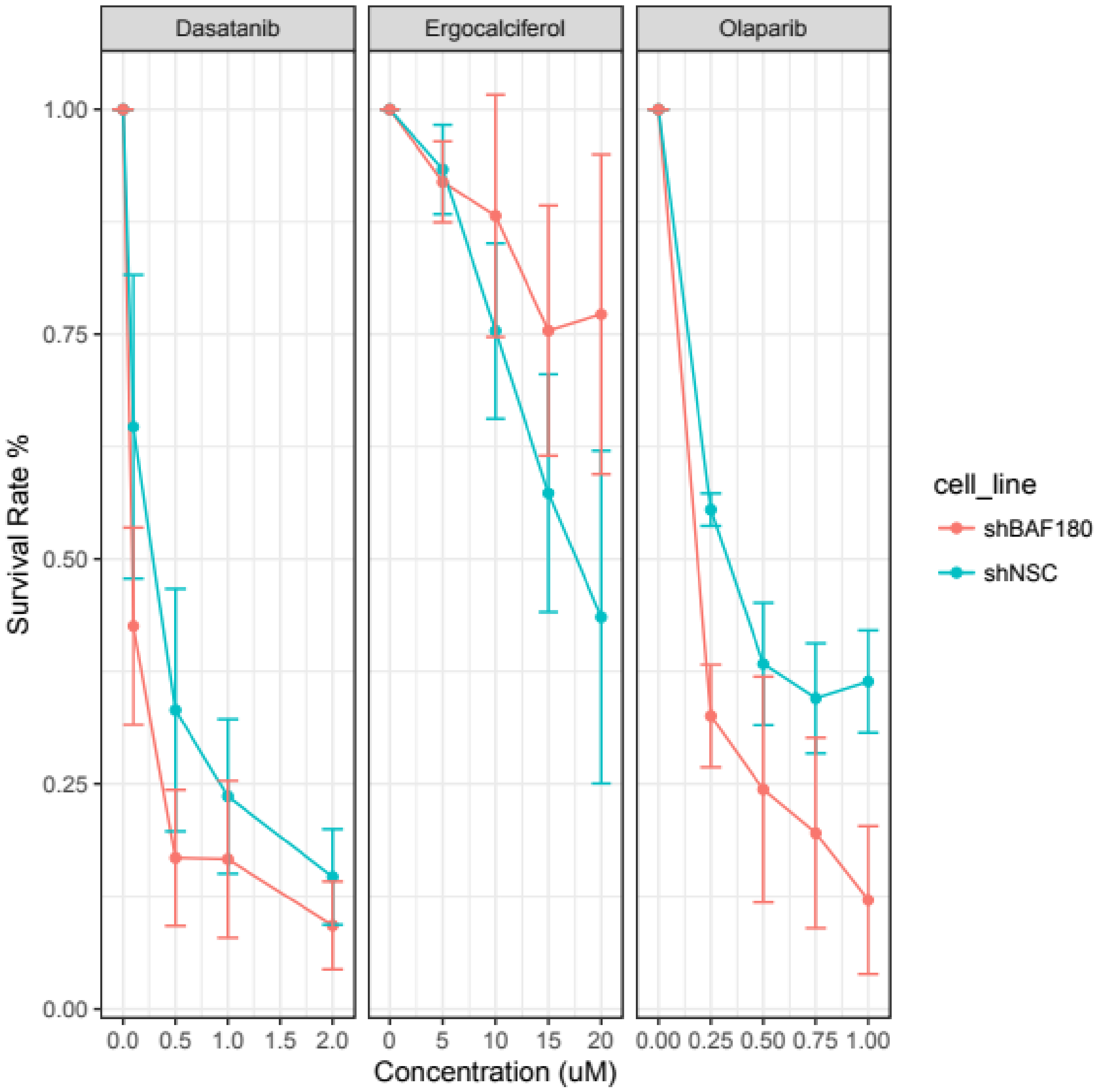


Figure 3.5. Carcinogenic survival assay results charting survival of PBRM1 / BAF180 knock-out cell lines with concentration intervals of the PARP inhibitor Olaparib, the POLA inhibitor Erocalciferol and the ABL inhibitor Dasatanib. These results suggest PBRM1 mutant cells may be more sensitive to both the PARP and ABL1 inhibitors while gaining some resistance to POLA1 inhibition. Error bars measure standard error of measurement. All drug intervals are measured in mM.

Of the drugs tested, three showed differential effects on the BAF180-deficient cells when compared to the control cells. PBRM1 mutant cells were more sensitive to both the PARP inhibitor and, to a lesser extent, ABL1 inhibitor than the control cells (Figure 3.5 with plate photography in supplementary Figure 3.5), whereas the PBRM1 mutant cells appeared less sensitive to the POLA1 inhibitor than the control cells (Figure 3.5). Interestingly, cells lacking ARID1A, which is another SWI/SNF subunit, are also selectively sensitive to PARP inhibitors (Geng et al., 2016; Shen et al., 2015), which supports this relationship. We also note this ARID1A / PARP1 SSL interaction was not present in the BioGRID data used to generate our training set but was also predicted with a high probability by Slant. The two protein products of the two genes SSL with PBRM1; PARP1 and ABL1, share a number of similar cellular processes such as regulation of differentiation, proliferation and of DNA damage and stress response. POLA1 which potentially shares a different type of interaction, synthetic rescue, plays an essential role in the initiation of DNA replication.

3.4 Discussion

In this paper we have predicted SSL relationships using features derived from both in-species and cross-species PPI network information. The SLant consensus classifier out-performs previous attempts at predicting human and model organism SSL interactions and may provide a useful tool in guiding future experimental validation of SSL pairs.

The original intention in this study was to predict cross-species without using the target species' data in the training set. However our in-species predictions generally performed so well it seemed sensible to instead use the additional cross-species data as an enhancement instead. The only in-species classifier that underperformed was that derived for *S. cerevisiae*. However, this result should be interpreted with caution; direct comparison of results is not possible as there are differences in the validation data. So that others may compare their algorithm to ours we have made all of the source code for SLant freely available so that our results, training data and validation can easily be recreated and repeated.

Improving the quantity and the quality of the input data will also improve the quality of the SSL and SDL predictions. For instance the amount of genetic interaction data is very limited in humans and *D. melanogaster*. Protein-protein interaction data is plentiful for humans and the model organisms studied, but the majority of the interactions are unlabelled. Adding additional annotation to these interactions, e.g. the direction of an interaction, may improve predictions if enough labelled data were available. Also, both the PPI and the genetic interactions reported have ‘popularity bias’; genes and proteins of biological or medical interest are frequently studied and hence more interactions involving them are reported.

Recently Abdollahpouri et al. (Abdollahpouri et al., 2017) developed a flexible regularization-based framework which can be used to control for popularity. An adaptation of this method to enhance the coverage of less frequently reported genetic interactions, may help mitigate this bias. Furthermore, providing a reliability score for genetic interactions and only using the more reliable ones may be particularly important for *S .cerevisiae* where although there is a wealth of data, the number of false positives reported experimentally may be corrupting the prediction accuracy.

In an attempt to ascertain whether synthetic lethal interactions occurred within or between local clusters of genes in our physical network we applied a spin glass random walk to assign genes to distinct clustered communities separated by choke points across the graph. Analysis showed that the majority of SSL interactions occurred between these communities rather than within them. Based on the shorter distance between SSL genes and higher occurrence of adjacency presumably SSL genes are often at the peripheries of these communities. Further exploration of how SSL pairs are distributed between clustered communities such as these may shed further light on the node wise features of genetic interactions.

Although this study does not use orthology data directly we do note that our GO annotation features may in some way serve as a proxy for orthology data and this study could be also be expanded in the future through improved analysis of the relationship between GO terms and pairwise SSL pairs.

The identification of SSL interactions is a key step in expanding and improving targeted cancer therapy. The results presented here suggest that inhibition of PARP1 or of ABL protein kinase 1 may have therapeutic value in tumours lacking functional BAF180. The computational and experimental validation of our models performance presented in this study suggests that the predictions provided by SLant, all of which have been made publicly available, will be useful in guiding future SSL screening studies and ultimately in the continued goal of generating a more complete list of human SSL pairs.

4 - Slorth: Validated and predicted synthetic lethal gene pairs with associated drug, disease and orthology data

4.1 Introduction

4.1.1 Synthetic lethal interactions may make suitable cancer drug targets

Synthetic lethal (SL) interactions are negative genetic interactions that can broadly be classified into two sub-types; synthetic sick lethal (SSL) and synthetic dosage lethal (SDL). Two genes are said to share a synthetic sick lethal (SSL) relationship when a disruption to either gene individually leaves the host cell viable while the deactivation of both genes simultaneously leads to cell sickness or death. Similarly two genes are said to be synthetic dosage lethal (SDL) when the disruption of one gene paired with an up-regulation in the other causes cell death (Hartwell et al., 1997).

SSL interactions can be exploited therapeutically for instance, when one gene, a tumour suppressor, is inactivated by a cancer driving mutation while the protein product of the other gene in the SSL pair is targeted and inactivated pharmacologically (Megchelenbrink et al., 2015). The most prominent cancer drugs in the clinic that exploit SSL interactions are the PARP inhibitors which are effective in a variety of BRCA1 and BRCA2 deficient cancers (Bryant et al., 2005; Tangutoori et al., 2015). Theoretically, SDL interactions could also be exploited to target tumours where an hard-to-target over-expressed oncogene shares an SDL interaction with a gene that has a therapeutically tractable protein product (Chan and Giaccia, 2011; Kaelin Jr and Kaelin, 2005).

Due to the experimental burden of exhaustive screening for human SSL interactions a large number of potentially therapeutically actionable interactions remain undiscovered (You et al., 2010). Computational methods for inferring and predicting SSL interactions are proving to be a promising way to guide experimental screening to mitigate this challenge (Benstead-Hume et al., 2017a) with published algorithms including (Jacunski et al., 2015; Jerby-Arnon et al., 2014; Kranthi et al., 2013; Li et al., 2011; Madhukar et al., 2015b; Paladugu et al., 2008; Ryan et al., 2014; Wang and Simon, 2013; Wong et al., 2004b; Wu et al., 2014, 2013; Zhang et al., 2015). We have recently developed, and experimentally validated, the Slant algorithm which predicts humans SSL interactions with a ROC AUC of 0.985, an improved performance on previously available algorithms (Benstead-Hume et al., 2019).

4.1.2 Existing synthetic lethal interaction databases

Considering the therapeutic potential of genetic interactions there are relatively few on-line resources available for easy browsing and discovery of clinically relevant synthetic lethal gene pairs.

BioGRID (Stark, 2006) is the prominent curated database that features a wealth of experimentally determined physical and genetic interactions for a range of species. BioGRID (version 3.5.172) features 598,168 negative genetic interactions including those labelled as ‘Synthetic lethality’, 'Synthetic Growth Defect', 'Dosage Growth Defect', 'Dosage Lethality' and the general ‘Negative Genetic’. These include 4,778 human negative genetic interactions, 24 fruit fly, 1,244 worm, 462,432 budding yeast and 39,799 fission yeast published across 4,230 publications. While search functionality is available, the on-line browsing and discovery ability is somewhat limited without downloading data and manual analysis.

SynlethDB (Guo et al., 2015) comprises of ~34,000 SL pairs. The data was derived from experimentally determined genetic interactions extracted from databases including Syn-Lethality (Li et al., 2014) GenomeRNAi (Gilsdorf et al., 2009) and BioGRID (Chatr-Aryamontri et al., 2015). The data were augmented by interactions extracted from manual curation of literature and text mining results, bi-specific shRNA screenings (Firth et al., 2009) and SL pairs for humans computationally predicted by the Daisy algorithm (which achieved a AUC ROC score of 0.779) (Ryan et al., 2014). Drug and orthologue data are available via SynlethDB although this functionality is not integrated into the application’s search tools. SynlethDB has not been updated since 2017 and does not appear maintained since 2018.

4.1.3 Slorth

Here we present the Slorth (Synthetic Lethality and ORTHology) database (http://slorth.biochem.sussex.ac.uk) database that has been designed to integrate 852,609 high quality SSL interaction predictions from the SLant (Synthetic Lethal analysis via network topology) algorothim (Benstead-Hume et al., 2019) with 331,308 experimentally determined negative genetic interactions for humans, worms, fruit fly, budding yeast and fission yeast sourced via the BioGRID database.

Slorth then combines these data with cancer, drug, pathways and orthologue information with the aim to enable researchers and clinicians to identify SL interactions with therapeutic potential.

Tools are available to quickly highlight interactions that feature genes associated with cancer, gene's whose protein products are possible drug targets, and the pathways that the proteins are involved in. Slorth's network visualizations highlight high-quality cross-species interactions and provide a wider view of the network of SL interactions surrounding a gene.

4.2 Methods and results

4.2.1 SLant

A brief summary of the Slant algorithm is described here. The SLant algorithm employs supervised machine learning classifiers to predict human and other model organisms synthetic lethal interactions by exploiting conserved patterns in biological network data both within and across model species.

SLant's network models were built using protein-protein interaction data from the STRING (von Mering et al., 2005) database for humans and each model organism. In these models nodes represent proteins and edges represent an interaction between two proteins. Network analysis algorithms were performed upon each network to extract a number of topological parameters be used as features used in classifiers trained to classify and predict SSL interactions.

The SLant algorithm employs a total of 17 network parameter features and 3 gene ontology (GO) term features. These network features are classified as either pairwise, a feature that relates to a gene pair or node-wise, a feature associated with an individual gene. Node-wise features include parameters such as degree, the number of edges entering or exiting a node and betweeness, the number of shortest paths across all nodes in the network that pass through a certain node. Pairwise topological features include parameters such as the shortest path between two genes and adhesion, the number of edges you would have to remove from a network to create two separate sub-networks separating the two nodes. Spinglass random walks were also performed on the biological networks to find natural clustering on proteins and their interactions, this community clustering provided further pairwise features with good predictive power. SLants GO term features were generated by integrating GO consortium (Ashburner et al., 2000) data with STRING data to model how GO terms are distributed across the network of protein interactions. A full list of features used in Slant is available in Supplementary Table 4.1.

SLant’s random forest classifiers are trained using these features for a large number of potential gene pairs which are labelled as either SSL or non-SSL, based directly from BioGRID data. SLant provides both in-species classification, where training and test data from a single organism is used for classification and validation and cross-species classification where models built using the training data for each model organism is used to to blindly predict SSL for each other species. Slant uses the results from the in-species classifier and each of the cross-species classifiers as features for a final consensus random forest classifier. The prediction score provided in the Slorth database give the proportion of trees in the consensus classifier that classified the interaction as SSL.

4.2.2 SLant validation

5-fold cross validation was used to to optimise parameters for each organism's random forest classifier and evaluate in-species classification performance. Data was segmented into training and test sets with 20% set aside for validation. Ultimately SLant’s consensus classifier achieved a mean AUC ROC of 0.985 (se. 0.008, n=10) in validation for the classification of human SL interactions out-performing all synthetic lethal classification systems previously published. budding yeast achieved an AUC ROC of 0.907, worm 0.982, fruit fly 0.903, and fission yeast 0.920 in validation for their respective consensus classifiers.

In addition, clonogenic survival assays (Franken et al., 2006) were used to perform experimental validation of a small sample of SLant human SSL predicted pairs with therapeutic potential. Using PBRm1 (BAF180) knockouts from the U2OS cell line and the selected inhibitors , some evidence was found to support interactions between 3 of 7 gene pairs screened including PBRm1 – PARP1, PBRm1 – ABL and PBRm1 – POLA; all gene pair predictions not found in the oringal SLant training data. Though this work is not conclusive it does further support the quality of the classification and predictive power of SLant.

4.2.3 BioGRID data

The SLant training set is labelled based on BioGRID data, specifically those genetic interactions labelled as either synthetic lethal or negative genetic or Synthetic Dosage Lethal. As such SLant's predictions do not include those interactions featured in BioGRID (Stark, 2006). To provide a more complete set of SL interactions the BioGRID data is included alongside the Slorth dataset with each interactions scored based on the number of reference papers available in BioGRID for that interaction.

4.2.4 Slorth Statistics

At time of writing Slorth includes 243,750 human SSL interactions featuring 4,474 genes . These genes are associated with 449 related drugs, 318 diseases and 3,403 GO terms. Slorth features a further 386,309 interactions for Budding yeast, 37,113 interactions for worm, 103,258 interactions for fruit fly and 42,211 interactions for fission yeast along with links to associated gene orthologs across species where available.

4.2.5 Functional annotation

Slorth's predicted and validated SL data has been integrated with cancer, drug, orthologue and pathway data to help users more easily browse and search for clinically relevant interactions.

Cancer mutation data was sourced via the COSMIC Cancer Gene Census (Forbes et al., 2016)(Futreal et al., 2004) and associated with 719 genes in the Slorth data where mutations have been causally implicated in cancer. Drug and inhibitor data were sourced through DrugBank (Wishart, 2006) which curates drug/target data from a range of sources including PubChem (Wang et al., 2009), KEGG (Ogata et al., 1999), PubMed (Canese and Weis, 2013), ChEBI (Degtyarenko et al., 2008), MetaCyc (Karp, 2002) and OMMBID (Welsh MJ, Ramsey BW, Accurso F, 2001). Orthologue data linking genes and interactions between species was sourced via Ensembl BioMart (Kinsella et al., 2011). Pathway data was sourced via GO annotation via the Gene Ontology consortium (Ashburner et al., 2000), a large scale bioinformatics project that curates experimental data from over 100,000 peer-reviewed papers to "annotate" gene function (Figure 4.1).

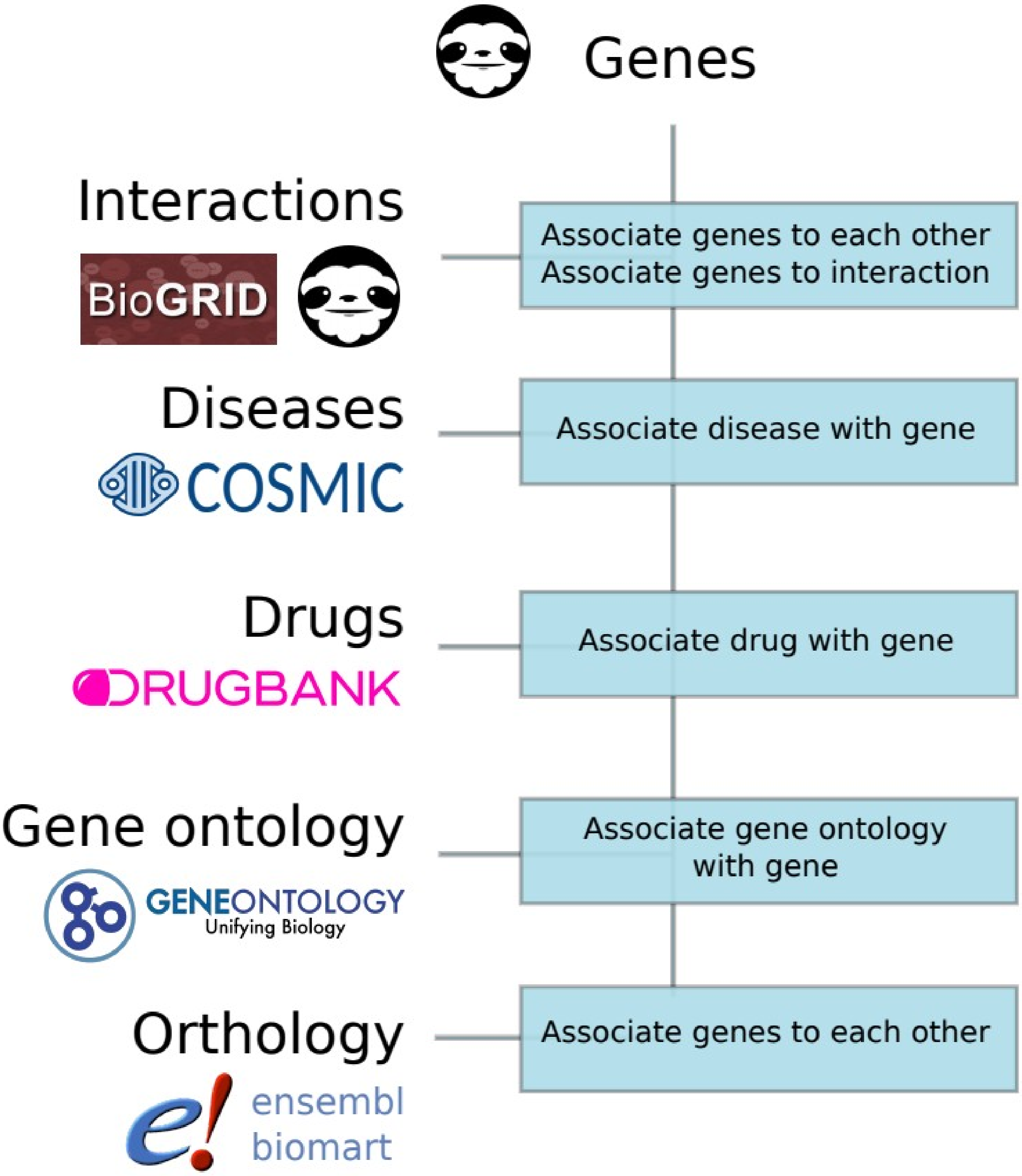

Genes
Interactions
BioGRID
Associate genes to each other
Associate genes to interaction
Diseases
COSMIC
Associate disease with gene
Drugs
DRUGBANK
Associate drug with gene
Gene ontology
GENEONTOLOGY
Unifying Biology
Associate gene ontology with gene
Orthology
ensembl
biomart
Associate genes to each other

Figure 4.1 Slorth database population workflow – A schematic describing the process taken to populate Slorth. Interaction data from Slorth and BioGRID was used to create relational associations between individual gene objects in a Ruby on Rails relational database. Interaction objected were also created and associated with their respective gene objects in many to many relationships. Disease data from cosmic, drug data from drugbank and gene ontology data from the gene ontology consortium were added as items to the database and associated with any associated genes. Orthology data was used to create ontology relationships between relevant gene pairs.

4.2.6 Slorth development and database population workflow

Slorth was built using the Ruby on Rails (version 4.2.8) web development platform (Bächle and Kirchberg, 2007).

To populate the Slorth database we use a Ruby on Rails helper script that reads raw CSV data files and creates rails active record objects for each row which are in turn passed to their respective table in a relational database. The Slorth database features tables for genes, interactions, diseases, drugs, orthology and GO ontology with data sourced from the resources discussed above.

After populating the gene table with gene names, identities and organism names for all genes with available interactions we populated our interaction table which includes the two interacting gene names, an interaction type, source and a score. Based on the information given by each interaction a many to many relational link was created between the two respective gene objects. Additional relationships were also added between each gene and their respective interaction object.

Once the database model for interactions and genes was complete we imported further data for diseases, drugs, gene ontology using the same helper script. In each case, for each row, a data object was created in the relevant table and a many to many relationship was created between the given data object and the related gene. Finally we imported orthology data. For each row of orthology data we create an relational link between the two corresponding genes.

4.3 Using Slorth

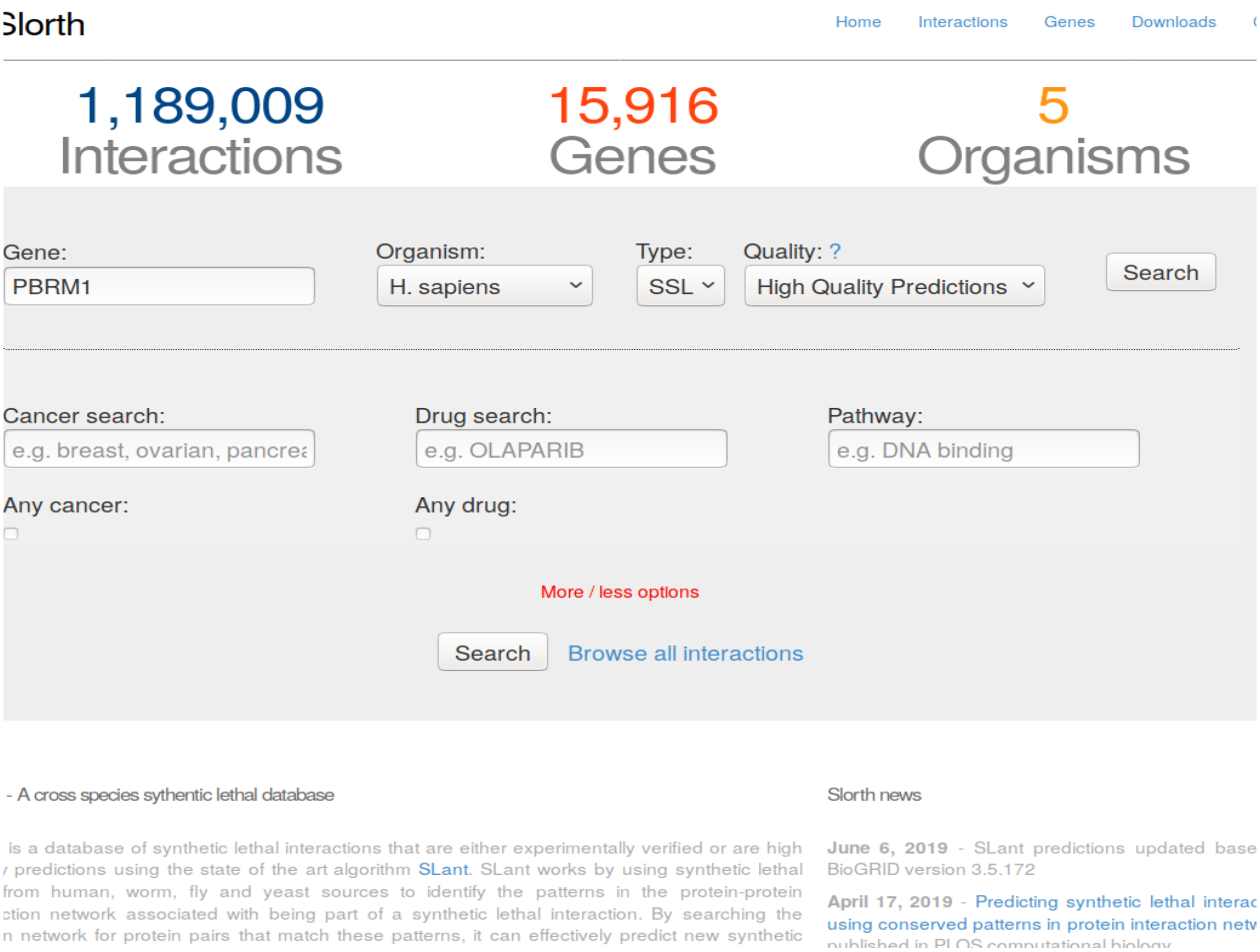
PBRM1 is commonly mutated in both clear cell renal carcinoma and breast cancers. As a tumour suppressor PBRM1 does not traditionally represent a suitable drug target but as part of a SSL pair it may still provide therapeutic opportunities.

We can search for a gene such as PBRM1 in Slorth at either the gene or interaction level. Interactions can also be filtered by drugs, diseases and pathways associated with either gene in each pair .

Performing a broad search for interactions associated with PBRM1 as in Figure 4.2 we are given a list of associated interactions as well as a network visualisation to help us understand the relationships between interactions.

The first two results are experimentally validated SSL interactions sourced via BioGRID. Interactions sourced from BioGRID are scored based directly on the number of references available for that interaction. We can see that these two results are both supported by one paper each. Links to these references are available on each interaction page.

Below these experimentally validated results are an additional 27 high quality predictions sourced via the SLant classifier. The first of these, PARP1 / PBRM1 reports a score of ~0.86. Interactions predicted via SLant are scored based on a random forest classification probabilistic output ranging from 0.5 which means the interaction was neither predicted by SLant as SL or Non-SL with any confidence to 1 which means the interaction was classified with a high confidence.



Slorth

Home Interactions Genes Downloads

1,189,009 Interactions

15,916 Genes

5 Organisms

Gene: PBRM1

Organism: H. sapiens

Type: SSL

Quality: ? High Quality Predictions

Search

Cancer search: e.g. breast, ovarian, pancre

Drug search: e.g. OLAPARIB

Pathway: e.g. DNA binding

Any cancer: □

Any drug: □

More / less options

Search Browse all interactions

- A cross species sythentic lethal database

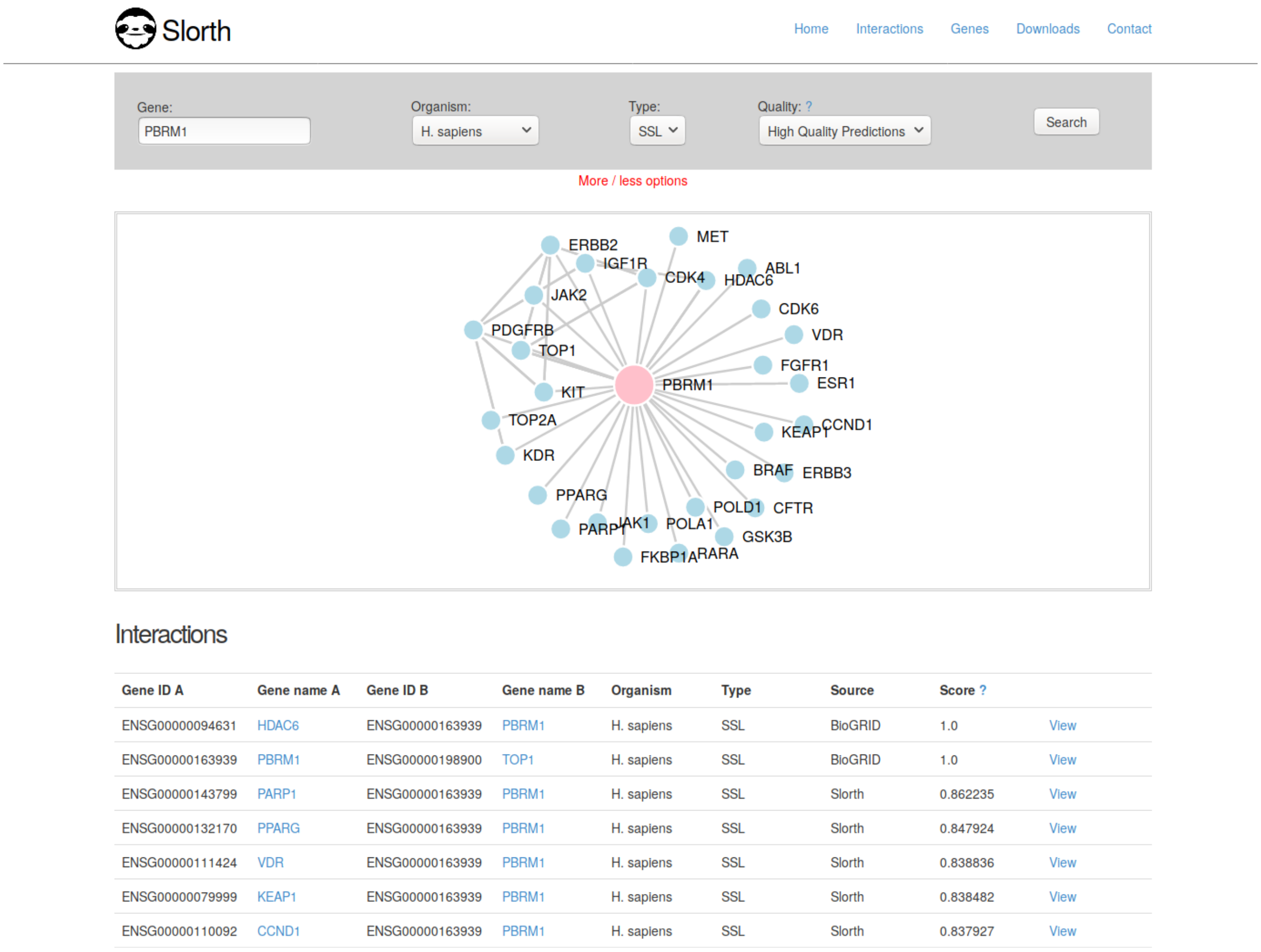
is a database of synthetic lethal interactions that are either experimentally verified or are high predictions using the state of the art algorithm SLant. SLant works by using synthetic lethal from human, worm, fly and yeast sources to identify the patterns in the protein-protein ction network associated with being part of a synthetic lethal interaction. By searching the n network for protein pairs that match these patterns, it can effectively predict new synthetic

Slorth news

June 6, 2019 - SLant predictions updated base BioGRID version 3.5.172

April 17, 2019 - Predicting synthetic lethal interac using conserved patterns in protein interaction netw published in PLOS computational biology

Figure 4.2 a.



Gene: PBRM1

Organism: H. sapiens

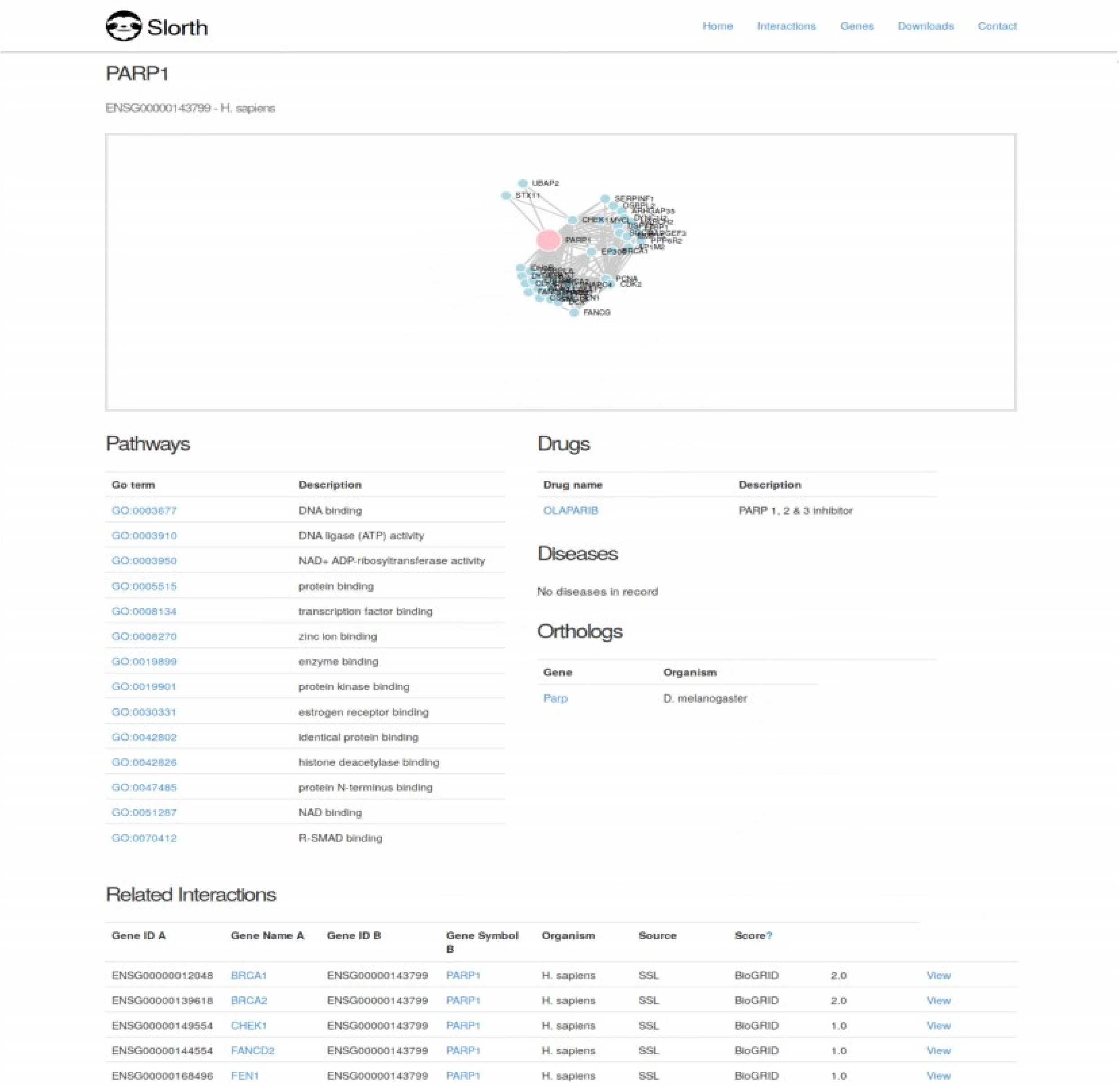
Type: SSL

Quality: ? High Quality Predictions

Interactions

Gene ID A	Gene name A	Gene ID B	Gene name B	Organism	Type	Source	Score ?	
ENSG00000094631	HDAC6	ENSG00000163939	PBRM1	H. sapiens	SSL	BioGRID	1.0	View
ENSG00000163939	PBRM1	ENSG00000198900	TOP1	H. sapiens	SSL	BioGRID	1.0	View
ENSG00000143799	PARP1	ENSG00000163939	PBRM1	H. sapiens	SSL	Slorth	0.862235	View
ENSG00000132170	PPARG	ENSG00000163939	PBRM1	H. sapiens	SSL	Slorth	0.847924	View
ENSG00000111424	VDR	ENSG00000163939	PBRM1	H. sapiens	SSL	Slorth	0.838836	View
ENSG00000079999	KEAP1	ENSG00000163939	PBRM1	H. sapiens	SSL	Slorth	0.838482	View
ENSG00000110092	CCND1	ENSG00000163939	PBRM1	H. sapiens	SSL	Slorth	0.837927	View

Figure 4.2 b.



PARP1

ENSG00000143799 - H. sapiens

Pathways

Go term	Description
GO:0003677	DNA binding
GO:0003910	DNA ligase (ATP) activity
GO:0003950	NAD+ ADP-ribosyltransferase activity
GO:0005515	protein binding
GO:0008134	transcription factor binding
GO:0008270	zinc ion binding
GO:0019899	enzyme binding
GO:0019901	protein kinase binding
GO:0030331	estrogen receptor binding
GO:0042802	identical protein binding
GO:0042826	histone deacetylase binding
GO:0047485	protein N-terminus binding
GO:0051287	NAD binding
GO:0070412	R-SMAD binding

Drugs

Drug name	Description
OLAPARIB	PARP 1, 2 & 3 inhibitor

Diseases

No diseases in record

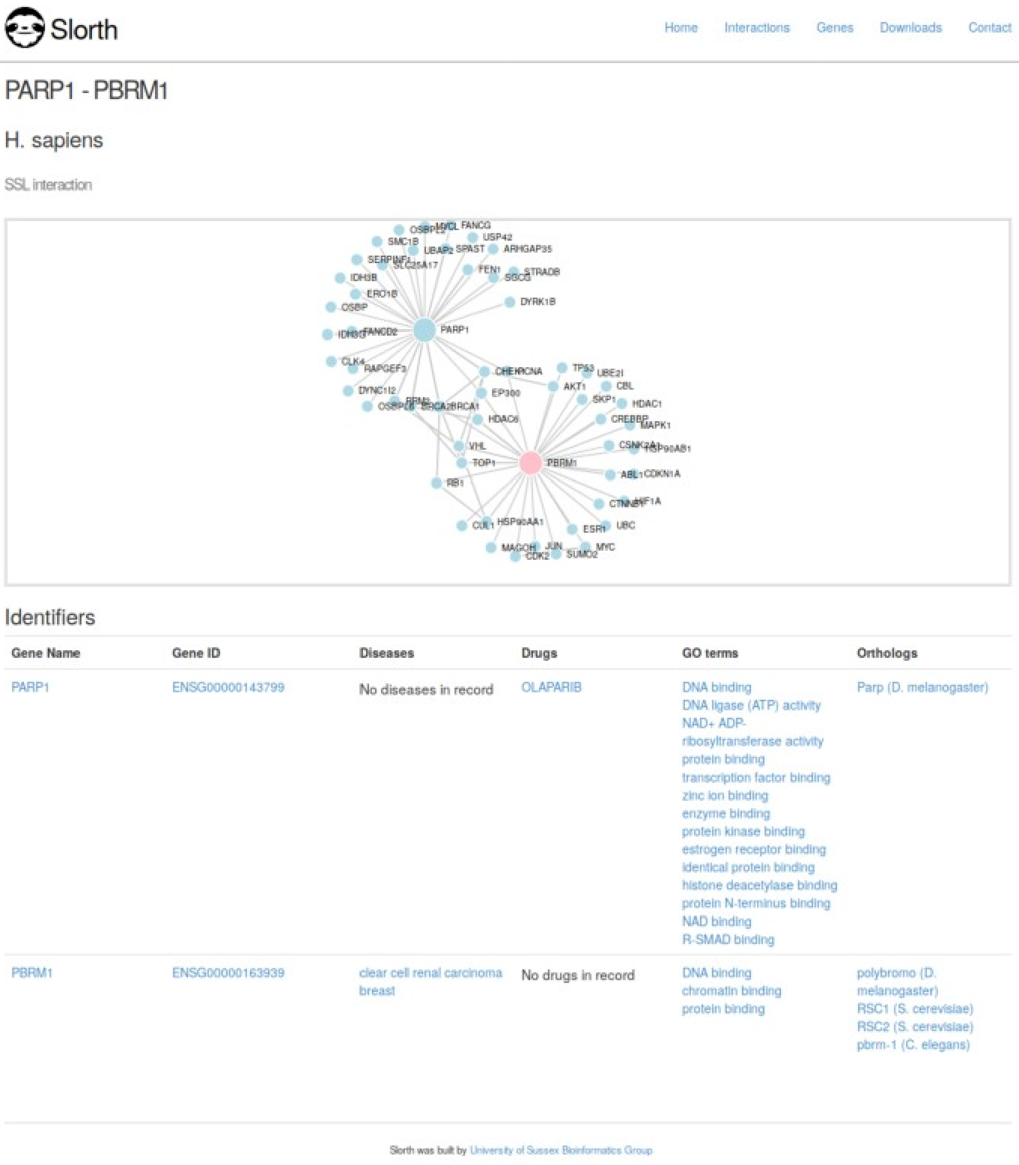
Orthologs

Gene	Organism
Parp	D. melanogaster

Related Interactions

Gene ID A	Gene Name A	Gene ID B	Gene Symbol B	Organism	Source	Score?		
ENSG00000012048	BRCA1	ENSG00000143799	PARP1	H. sapiens	SSL	BioGRID	2.0	View
ENSG00000139618	BRCA2	ENSG00000143799	PARP1	H. sapiens	SSL	BioGRID	2.0	View
ENSG00000149554	CHEK1	ENSG00000143799	PARP1	H. sapiens	SSL	BioGRID	1.0	View
ENSG00000144554	FANCD2	ENSG00000143799	PARP1	H. sapiens	SSL	BioGRID	1.0	View
ENSG00000168496	FEN1	ENSG00000143799	PARP1	H. sapiens	SSL	BioGRID	1.0	View

Figure 4.2 c.



PARP1 - PBRM1

H. sapiens

SSL interaction

Identifiers

Gene Name	Gene ID	Diseases	Drugs	GO terms	Orthologs
PARP1	ENSG00000143799	No diseases in record	OLAPARIB	DNA binding DNA ligase (ATP) activity NAD+ ADP-ribosyltransferase activity protein binding transcription factor binding zinc ion binding enzyme binding protein kinase binding estrogen receptor binding identical protein binding histone deacetylase binding protein N-terminus binding NAD binding R-SMAD binding	Parp (D. melanogaster)
PBRM1	ENSG00000163939	clear cell renal carcinoma breast	No drugs in record	DNA binding chromatin binding protein binding	polybromo (D. melanogaster) RSC1 (S. cerevisiae) RSC2 (S. cerevisiae) pbrm-1 (C. elegans)

Slorth was built by University of Sussex Bioinformatics Group

Figure 4.2 d.

Figure 4.2. Work flow example for Slorth database – Example of interaction search page (a.) and results for 'PBRM1' filtered for high quality predictions results in 29 results including both experimentally valid results form BioGRID and high certainty predictions from Slant (b.) . From the results page additional information such as associated diseases, drugs, pathways and orthology is available for both individual genes in a gene results pages (c.) and for interactions in an interaction results pages (d.).

To help researchers control the quality of interactions in their search Slorth provides quality filtering in its search suite allowing users to filter for experimentally validated entries only, high quality SLant predations, which includes validated entries and all entries. Experimentally validated interactions include only those featured in BioGRID. High quality predictions include both experimentally validated interactions and interactions predicted via SLant with a confidence score of above 0.75.

The interaction view page provides further detail on the genes associated with that interaction. Slorth provides network visualizations for both genes and interactions to provide a wider view of

the network of SSL interactions surrounding a gene. In these network visualisations each node denotes a gene and each edge an interaction. From this we can see that BRCA2, CHEK, and EP300 share SSL interactions with both PARP1 and PBRM1 all of which may present therapeutic opportunities.

Below the network visualisation a table contains additional data for both genes in the interacting pair. The GO term data provided shows that both genes are associated with DNA and protein binding and although SL pairs are not always conserved across species (Wu et al., 2013) the orthology section may help researchers relate gene pairs between organisms. In this case we see that PARP1 has a known ortholog in *D.*

melanogaster and *PBRM1* have orthologs in *S. cerevisiae, C. elegans* and *D. melanogaster.*

Finally we can see that PBRM1 is associated with clear cell renal carcinoma and breast cancer while PARP1 is associated with a drug, OLAPARIB, a PARP1 inhibitor. This result may guide the researcher to explore the potential of PARP1 as a therapeutic target in a PBRM1 deficient tumour..

4.4 Conclusion

Slorth features some of the highest quality SL predictions available via SLant along with a full compliment of experimentally validated SL pairs via BioGRID. Furthermore Slorth provides an interface that has been designed to enable easy discovery of clinically relevant pairs. The ultimate aim of Slorth is to help better guide future SL screening and ultimately further the development of targeted drug therapies to improve patient outcome and quality of life.

5 - Biological network topology features predict gene dependencies in cancer cell lines

5.1 Introduction

An essential gene is one which is necessary for cellular survival and reproductive success. However, the exact set of essential genes is context specific depending on the cell type, genetic and epigenetic aberrations and the cell environment. The different definitions and measurements of essentiality often have considerable overlap but there are also large areas of disagreement (Bartha et al., 2018; Eisenberg and Levanon, 2013).

During the process of carcinogenesis, the pattern of essential genes changes as cells become addicted to oncogenes and tumour suppressor genes become inactivated (Luo et al., 2009b; Weinstein, 2002). Identifying gene dependencies that result from carcinogenesis can provide opportunities for targeted treatments, as the inhibition of proteins which are essential in cancer cells but not in normal cells can lead to selective cell death (Workman et al., 2013). However, the heterozygous nature of cancer and the large number of genetic alterations in cancer cell lines prevent the exhaustive identification of these acquired essential proteins for all possible cell lines.

Several groups have used features derived from protein-protein interaction (PPI) networks to predict cancer genes (Li et al., 2009), and genetic interactions (Benstead-Hume et al., 2019). Furthermore there have been a number of successful attempts to predict common essential genes using biological network data in different contexts and in different organisms (for a review see Zhang et al. (Zhang et al., 2016)). These studies have used a range of different network data including protein-protein interaction (PPI) networks, transcriptional regulatory networks, gene co-expression networks, metabolic networks (Mns) and networks that integrate two or more of the above. Due to data availability these studies have generally focused on model organisms. For studies on *S. cerevisiae* see (Acencio et al., 2009; Chen and Xu, 2005; Saha and Heber, 2006). For studies on *E. coli* see (Hwang et al., 2009; da Silva et al., 2008) and for studies on various bacteria see (Cheng et al., 2014; Lu et al., 2014; Plaimas et al., 2010). For the most part these studies employ similar methods where topology data is extracted from the biological networks. This topology data is subsequently used as a feature set to train machine learning models to identify essential genes. For example, Saha et al. (Saha and Heber, 2006) reported a ROC AUC of 82% using PPI network degree count and conservation score features to classify ~2,200 essential genes in *S. cerevisiae* and Müller da Silva et al. (da Silva et al., 2008) who reported F1 scores of 83.4% for essential gene predictions and 79.7% for non-essential gene prediction in *E coli.* Similar levels of prediction have not been reported for human cell lines.

Generally past studies have focused on a static version of the known PPI network with little modification for individual samples. Observations made by Roumeliotis et al. (Roumeliotis et al., 2017), suggest that the effect of genetic variations can be transmitted from directly affected proteins to distant gene products through protein interaction pathways, suggest that the inclusion genetic alterations may allow us to improve the traditional PPI network model.

Recently there have been significant efforts to identify and catalogue cancer specific acquired essential genes, otherwise known as gene dependencies, experimentally. Amongst these efforts are a number of loss of function screens (Ngo et al., 2006) performed using both RNAi and CRISPR-Cas9 systems (Aguirre et al., 2016; Aksoy et al., 2014; Cheung et al., 2011; Luo et al., 2008; Marcotte et al., 2012, 2016). These screens investigate the changes in phenotype caused in cell lines by systematically knocking genes out one by one either through deletion or disruption. Knock-outs that result in significantly deleterious phenotypes signal that the respective gene may be essential in that cell line.

In response to reported off-target effects observed in loss of function screens, where genes other than the target are disrupted by certain RNAi (Aguirre et al., 2016; Birmingham et al., 2006; Buehler et al., 2012; Jackson and Linsley, 2004; Munoz et al., 2016), Tsherniak et al. (Tsherniak et al., 2017) building on previous work by Cowley et al. (Cowley et al., 2014), performed 285 genome scale systematic loss-of-function screens to identified cancer dependencies across a total of 501 human cancer cell lines covering 21 different tissue types. They found 6,476 genes that had a cancer dependency score of over 0.65 in at least one cell line. Of these 6,476 genes, 545 were dependencies in 20-50% of cell lines in at least one tissue-type. This suggested that these genes are commonly essential in cancer cells of that tissue type but non-essential in normal cells.

While identifying general essential genes or disease specific gene dependencies provides a better understanding of potential disease specific targets, loss of function screens are not readily available for the majority of individual cancer patients. Tools that could predict cell line specific gene dependencies from more readily available data such as mutations and gene expression may offer new opportunities for affordable tailored therapies (Benstead-Hume et al., 2017b; Charlton and Spicer, 2016).

In this study we use recent cell line specific gene dependency data along with PPI networks data to build models able to identify novel cell line specific gene dependencies. To do this we model genetic alterations in specific cell lines by perturbing their respective PPI networks. We explore the viability of identifying cell line specific gene dependencies both within and between various human cancer cell lines using this perturbed PPI networks data. Finally, we introduce DependANT, a classifier trained to predict cell line specific gene dependencies using both generic and perturbed PPI networks data with the aim of providing a low cost approach to identifying personalised cancer drug targets without the cost of exhaustive loss of function screening.

5.2 Methods

5.2.1.1 Constructing the base PPI

Our base protein-protein interaction data was obtained via the STRING database (v.10) (von Mering et al., 2005). This data was filtered to include only interactions with an experimental score higher that 80 to ensure each interaction was reliable. The ENSP protein IDs in this data set were converted to their respective ENSG gene IDs using Ensembl data (Hubbard et al., 2002). R (version 3.4.0) and the igraph package (version 1.1.2) (Csárdi and Nepusz, 2006) were used to produce a network model of the PPI data for each cell line.

5.2.1.2 Essentiality data and labelling

The Tsherniak et al. (Tsherniak et al., 2017) survival screen data, via project Achilles, provides a likelihood score for each gene in each cell line being a essentiality. We the same likelihood threshold as Tsherniak et al. to label each gene in our model as a gene dependency, those above 0.65 or non- dependency those below 0.65 for each cell line.

5.2.1.3 Perturbing the PPI

All edges in the directed PPI network have a weight of (0,1] which reflects the strength of expression of the initial protein, i.e. proteins that are not expressed have edges of weight 1 emanating from them, and as expression increases so the weight reduces. These weights are determined by modifying RNA seq data to reflect the loss and gain of function of proteins with mutated gene sequences.

In order to create these weights, RNA seq data from was downloaded from the Cancer Cell Line Encyclopaedia (Cancer and Line, 2015) , and mutation data was downloaded from Tsherniak via Achilles (Tsherniak et al., 2017).

Mutations that lead to loss and gain of function were identified or predicted as follows. Frameshift indels were assumed to lead to loss of function. The program SIFT (Kumar et al., 2009) was then used to remove mutation which were predicted to have no functional impact. The remaining missense mutations were categorised as leading to either loss of function or gain of function using a version of the MOKCARF algorithm (Baeissa et al., 2017b). MOKCARF uses features from Mutation Assessor (Reva, B.A., Antipin, Y.A. and Sander, 2010), Polyphen2 (Adzhubei et al., 2013) and FATHMM (Shihab et al., 2013) as input to a ADA boost classifier which has been trained on protein domains mutated in proto-oncogenes, or tumour suppressor to predict loss or gain of function.

Gain of function is assumed to have a multiplicative impact on RNA expression (set here to a factor of 10), whilst loss of function sets the resulting weight to 1.

Equation 5.1

$$weight(p)=max(0.5(1-\tanh(\ln(ex(p)*gof(p)))),lof(p))$$

where p is the protein and and ex(p) is the RNA_seq expression associated with protein p, $gof(p)=10$ if there is a mutation in the gene associated with protein p leading to gain of function, otherwise $gof(p)=1$.

and $lof(p)=1$ if there is a mutation in the gene associated with protein p leading to loss of function, otherwise $lof(p)=0$.

Tanh was used to constrain the resulting score to values between -1 and 1.

5.2.1.4 Feature generation

R and the igraph package were used to extract 14 network topology features for each cell line's protein interaction network described in Table 5.2.

5.2.1.5 Preprocessing feature data

To improve performance in cross cell line classification each cell line's feature set was normalised (Jacunski et al., 2015). To ensure unbiased validation we held-out 20% of this data to be used as a test set leaving 80% to be used as training data.

5.2.1.6 Model validation

Classification was performed using the R caret library's "ADA" boosted classification trees classifier. 5-fold cross validation was applied to each cell-lines training data to select the most optimised set of hyper-parameters. The ADA classifier as implemented in the caret library has three hyper-parameters to optimise, number of trees, max tree depth and learning rate.

A final model using these optimised hyper-parameters was then used to predict against the hold-out test set to assess predictive performance within each cell line and between each cell lines. These predictions were outputted as the probability of each class, essential or non-essential.

5.2.1.7 Pan-cancer model and unlabelled predictions

To predict dependency genes in unlabelled cell lines we first concatenated all training data into one large labelled training data set. We produced a number of feature sets for cell lines that were not included in the original training data and predicted dependency genes in these unlabelled cell lines based on a model trained on the pan cancer set.

5.2.1.8 Experimental validation

We chose a single unlabelled cell line, MCF7, for experimental validation. MCF7 was not featured in our training data and was chosen based on ready availability and good class balance for predictions on genes featured as part of the available DDR gene library.

We performed a high-throughput siRNA screen for experimental validation. Human breast (adenocarcinoma) MCF7 cells (validated by ATCC STR.V profiling) were grown in MEM supplemented with 10% FCS, penicillin/streptomycin and L-glutamine at 37oC and 5% CO2.

Cells were reverse transfected with library siRNA using lipofectamine RNAiMAX (as per the manufacturer’s instructions) in black 96 well plates. Plates were incubated at 37°C, 5% CO2 for 72 hours. CellTitre-Blue was added to determine cell viability, plates were analysed using a plate reader at 560/590nm.

5.2.1.9 Druggability annotation

Druggability annotation was performed using Cansar Black’s cancer protein annotation tools (Bulusu et al., 2014). We designated any genes with a “nearest drug target” score of 100% as a known drug target and any gene with one or more predicted drug targets in three dimensional structures that exhibited 100% homology with the respective gene’s sequence Identity.

5.3 Results

5.3.1 Data sets

DependANT classifies cell line specific gene dependencies via models built using protein-protein interaction (PPI) network and genetic alteration data. The PPI networks were sourced via STRING (von Mering et al., 2005) and the mutation and gene expression data used to perturb our networks, as well as the gene dependency scores used to label our training data, are publicly available from Tsherniak et al. via project Achilles (Tsherniak et al., 2017).

We selected all breast, kidney and pancreatic cancer cell lines that had sufficient gene dependency and genetic alteration data in the Tsherniak data (Supplementary Figure 5.1). These included 19 breast, 11 kidney and 11 pancreatic cell lines. For each cell line we selected all genes with a likelihood score higher than 0.65 in the Tsherniak study as a gene upon which its host cell is dependent, a total of 4,030 gene dependencies across 39 cell lines.

5.3.2 Gene dependency count and magnitude of genomic alteration are significantly correlated

We first set out to find out if and how acquired gene dependencies differ across cell lines and tissue types and how gene dependency is related to genomic alteration. Using the data sourced via Tsherniak et al. we first plotted the number of gene dependencies reported for each cell line against a measure of that cell line's genomic alteration.

We measured each cell line's level of genomic alteration by counting the number of genes that had pathogenic mutations as identified by SIFT (Sim et al., 2012) and the number of genes differentially expressed when compared to the mean expression level for cell lines of that tissue type, using a cut-off point of 0.5 TPM.

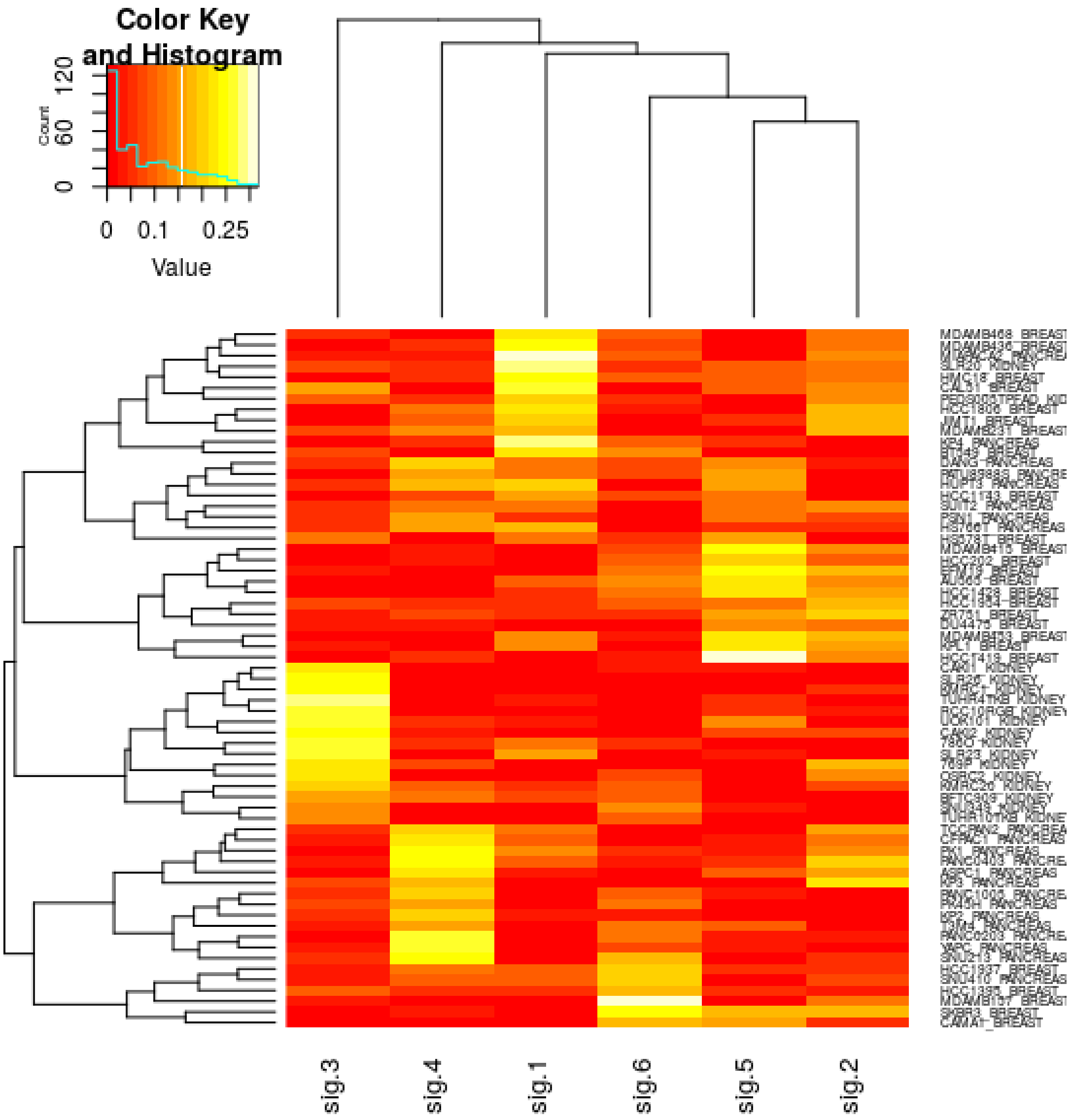
Across all cell lines we found a slight but significant positive correlation between the measure of genetic alteration and the number of gene dependencies in cell lines (R= 0.36, p=0.012) (Supplementary Figure 5.2. a.). To calculate the significance of this level of correlation we shuffled the data for genomic alterations 10,000 times, calculating the correlation coefficient each time to provide a normal distribution of correlation coefficients (Supplementary Figure 5.2. b.).

This significant positive correlation may be the result of alterations that have affected one or more otherwise non-essential genes that are part of synthetic lethal genetic interactions rendering the surviving gene in the pair as essential for cell viability.

We found that when compared to the other two tissue types cell lines originating in breast tissues exhibited, on average, a higher level of genomic alteration ($p=3*10^{-5}$) and a higher number of reported gene dependencies (p=0.024).

5.3.3 Gene dependency signatures are enriched for specific disease tissue types

In order to quantify how gene dependencies are distributed across specific tissue types we next performed non-negative matrix factorisation (NMF) in order to find common signatures of gene dependency. To better understand how these signatures relate across tissue types we added additional cell lines from pancreatic tissue samples. To render the data more easily manageable for NMF we filtered our gene dependency data to remove genes that showed low variation between tissue types, i.e. any genes with var <0.1 across all tissue types were removed from the data before factorisation.


Color Key
and Histogram
Count
0
60
120
0
0.1
0.25
Value
MDAMB468_BREAST
MDAMB436_BREAST
MIAPACA2_PANCREA
SLR20_KIDNEY
HMC18_BREAST
CAL51_BREAST
PEDS005TPEAD_KID
HCC1806_BREAST
JIMT1_BREAST
MDAMB231_BREAST
KP4_PANCREAS
BT549_BREAST
DANG_PANCREAS
PATU8988S_PANCRE
HUPT3_PANCREAS
HCC1143_BREAST
SUIT2_PANCREAS
PSN1_PANCREAS
HS766T_PANCREAS
HS578T_BREAST
MDAMB415_BREAST
HCC202_BREAST
EFM19_BREAST
AU565_BREAST
HCC1428_BREAST
HCC1954_BREAST
ZR751_BREAST
DU4475_BREAST
MDAMB453_BREAST
KPL1_BREAST
HCC1419_BREAST
CAKI1_KIDNEY
SLR26_KIDNEY
KMRC1_KIDNEY
TUHR14TKB_KIDNEY
RCC10RGB_KIDNEY
UOK101_KIDNEY
CAKI2_KIDNEY
786O_KIDNEY
SLR23_KIDNEY
769P_KIDNEY
OSRC2_KIDNEY
KMRC20_KIDNEY
BFTC909_KIDNEY
SNU349_KIDNEY
TUHR10TKB_KIDNE
TCCPAN2_PANCREA
CFPAC1_PANCREAS
PK1_PANCREAS
PANC0403_PANCRE
ASPC1_PANCREAS
KP3_PANCREAS
PANC1005_PANCRE
PK45H_PANCREAS
KP2_PANCREAS
T3M4_PANCREAS
PANC0203_PANCRE
YAPC_PANCREAS
SNU213_PANCREAS
HCC1937_BREAST
SNU410_PANCREAS
HCC1395_BREAST
MDAMB157_BREAST
SKBR3_BREAST
CAMA1_BREAST
sig.3
sig.4
sig.1
sig.6
sig.5
sig.2

]

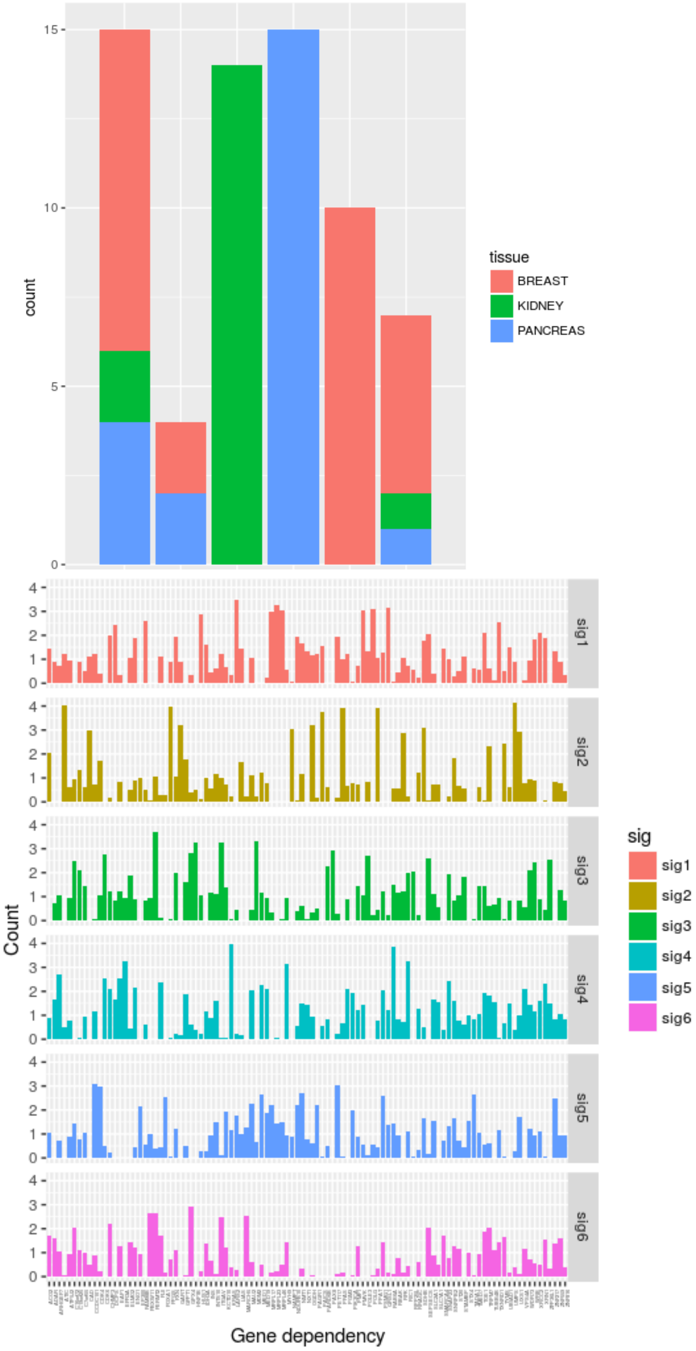


Figure 5.1 – Genedependency signatures derived from non-negative matrix factorisation

a. A clustered heatmap shows the clustering of gene dependency signature prominence across *cell lines. Dependency signature prominence sourced via the basis matrix (also known as matrix W) given by negative matrix factorisation.*

b. Enrichment analysis shows that tissue type is predictive of prominent gene dependency signature. Signature 6 for example is fully enriched for kidney cell lines, signature 2 for breast and signature 3 prominently features pancreatic cell lines.

c. The composition of each gene dependency signature given by the mixture coefficients matrix (or matrix H)

We found that six signatures was the minimum number required to describe the majority of the data (Figure 5.1. a.). We took the most representational signature for each cell line and called this the cell line's prominent signature. We plotted a count of cell lines with each corresponding prominent signature which was further grouped for tissue type to find enrichment. We found that two signatures contained only one type of tissue type, signature 2 which features only breast and signature 5 which features only kidney tissue. Signature 3 was also highly enriched for pancreatic tissue (Figure 5.1. b.).

This may suggest that different tissue types feature fairly stable, unique patterns of gene dependency either as a result of cellular environment or, especially in the case of cancer cell lines, synthetic lethal interactions.

Sig 1	Sig 2	Sig 3	Sig 4	Sig 5	Sig 6
GART	FOXA1	EFR3A	POLG	ELMO2	PAX8
ATIC	STX4	KRAS	MRPL23	GPX4	MDM2
CAD	MARCH5	TUBB4B	MRPL46	ITGAV	HNF1B
PAICS	TADA1	RAB6A	HUS1	VPS4A	RPP25L
PFAS	EP300	SLC7A1	LARS2	FERMT2	PARD6B
NAMPT	FBXW11	MYH9	MRPL17	SEPSECS	ZFP36L1
FPGS	CCDC101	ARHGEF7	QRSL1	MARCH5	POLE3
UMPS	PIK3CA	VPS4A	DCPS	CHMP3	CDK6
LIAS	CDK4	ADAR	PMVK	UBIAD1	FERMT2
OGDH	MED1	EAF1	TXNRD1	SMARCA4	C16orf72

Table 5.1. Prominently differentiated genes between gene dependency signatures. For each signature every gene was ranked by distance from the mean score given by the basis matrix compared to the same gene across all other signatures.

We generated a list of the most prominently differentiated genes (Figure 5.1. c.) by ranking the distance of each gene's occurrence count in each signature from the mean number of occurrences of that gene across all signatures as reported in Table 5.1.

5.3.4 Modelling cell lines with biological network and genetic alteration data

For each selected cell line a model was created from the STRING PPI networks data (von Mering et al., 2005). In each model a node represents a protein and each edge between nodes a physical interaction between the two respective proteins. Once each model is generated in this way we essentially treat each node as the gene associated with the protein (Figure 5.2).

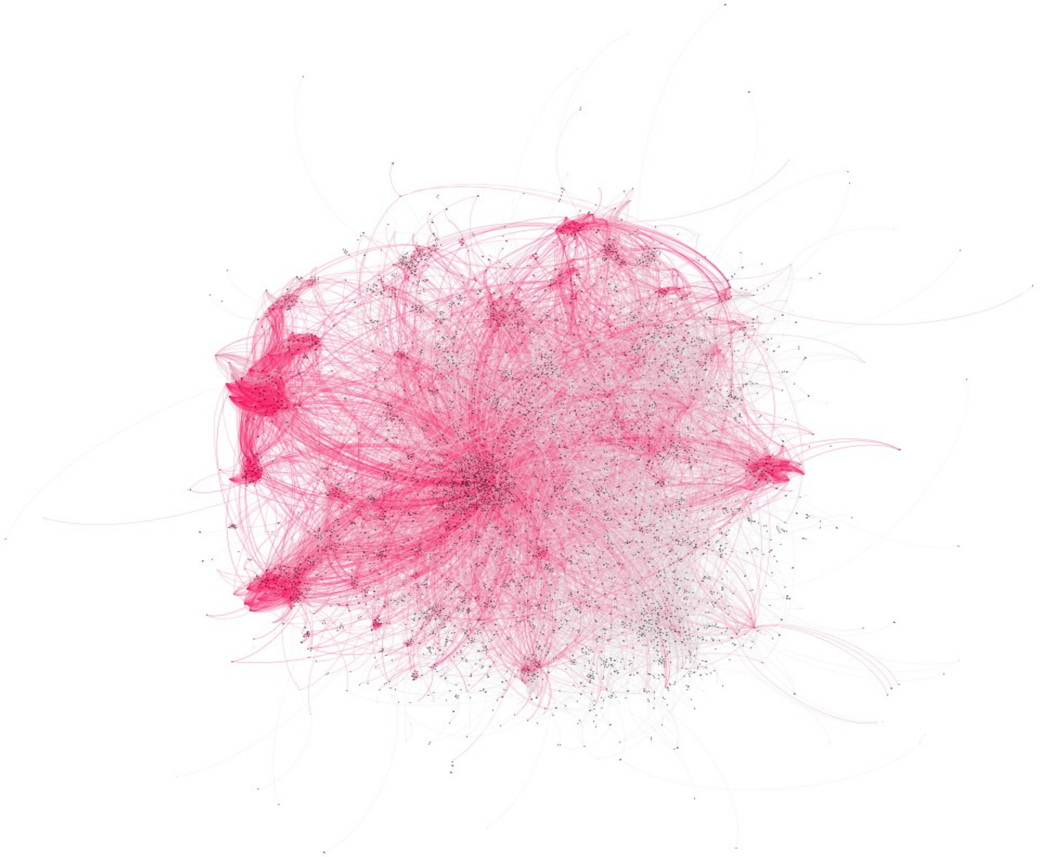


Figure 5.2. Plot of the PPI network graphs for breast cell line AU565 BREAST highlighting acquired essential genes in red suggests clustering of these gene dependencies.

We then extracted topology data for each node (Table 5.2) and used these data points as features in our machine learning models. The distribution of features values for dependency genes are somewhat different to those of non-dependency genes notably for the betweenness, constraint, eigen centrality and hub_score features (Supplementary Figure 5.3) suggesting that these features should provide some predictive power.

Feature name	**Description**
Betweenness	The number of shortest paths in the entire graph that pass through the node.
Constraint	Related to ego networks. A measure of how much a node's connections are focused on single cluster of neighbours.
Closeness	The number of steps required to reach all other nodes from a given node.
Coreness	Whether a node is part of the k-core of the full graph, the k-core being a maximal sub-graph in which each node has at least degree k.
Degree	The number of edges coming in to or out of the node.
Eccentricity	The shortest path distance from the node farthest from the given node.
Eigen centrality	A measure of how well connected a given node is to other well-connected nodes.
Hub score	Related to the concepts of hubs and authorities the hub score is a measure of how many well linked hubs the node is linked to.
Neighbourhood n size	The number of nodes within n steps of a given node for n of 1, 2, 5 and 6.

Table 5.2. List of graph topology features extracted from protein interaction network data with descriptions

For training purposes we labelled the nodes in PPI network using the gene dependency data sourced via Tsherniak et al. (Tsherniak et al., 2017) for each cell line as either a dependency or non-dependency. We refer to this unperturbed labelled PPI model as our base PPI networks model.

5.3.5 Base PPI network parameter data predicts pan-cell line dependency genes

To establish baseline performance for our classification models and to generate a list of relatively common dependency genes across cell lines we ran our classifiers on each cell line with no alterations or perturbations using the base PPI network discussed above.

We ran these classifiers to validate performance within cell lines, across cell lines of the same tissue type and across cell lines originating from different tissue types to understand how well the classifiers generalise.

To validate classification within individual cell lines we optimised our ADA boost classifiers' hyper parameters using 5-fold cross-validation on our training data and further validated the classification performance using hold-out test data which constituted 20% of the full data set.

We validated the model separately on each of our 42 cell lines, using both training data and validation data extracted from the same single cell line. Each trial was repeated 10 times using the base PPI model. This gave us a mean predictive performance of AUC ROC 0.765 (s.d. 0.024).

To measure performance across cell lines originating from the same tissue type and the predictive performance between tissue types we used the training sets that were already generated for each cell line to train our classifiers and we systematically validated each cell line against each other cell lines test set.

To ensure that our models were not being biased by genes that were present in both training and test sets we ensured that any genes present in the training set were removed from the active test set.

We first measured how well our models generalise from one cell line to another within the same tissue type. Under these conditions the base PPI models had an average AUC ROC of 0.761 (s.d. 0.005), 0.755 (s.d. 0.008) and 0.754 (s.d. 0.012) for breast, kidney and pancreatic cell line sets respectively.

Finally, we trained our model on kidney data before predicting acquired gene dependencies in breast and pancreatic tissue. These cross cell line predictions resulted in a mean AUC ROC of 0.758 (s.d. 0.007) and 0.758 (s.d. 0.01) respectively. Similarly when we trained the model on breast data before predicting dependency genes in kidney tissue the model had a mean AUC ROC of 0.759 (s.d. 0.006) and breast to pancreas performed similarly with 0.761 (s.d. 0.01) Taking the mean performance of all cell lines predicting all other cell lines the base PPI network model gave an AUC ROC of 0.757 (s.d. 0.007).

5.3.6 Feature importance

To quantify which features provide the most predictive power to our models we calculated a normalised importance score for each feature for each cell line and took the distribution of these scores across all cell lines. Feature importance was calculated by measuring the mean decrease in accuracy holding out each feature across all tree permutations in a random forest.

We found that a number of features that measure connectivity of a gene perform better than degree centrality although degree centrality does provide a moderate amount of predictive power. PageRank and eigen centrality scored well in all cell line models followed by hub score and constraint. Eccentricity, the distance a given node is away from the furthest node from itself in the network, a measure of how close that node is to the centre of the network, performs badly across all models.

These importance scores reflected the class feature distributions fairly well, i.e. features whose values varied more between essential and non-essential genes provided more predictive power. PageRank and constraint showed a noticeable differentiation between classes whilst the differentiation between classes for eigen centrality and hubscore features were not as prominent (Supplementary Figure 5.4).

5.3.7 Our perturbed models reported improved predictive power compared to our base model

Our base PPI models performed moderately well when predicting commonly observed essential genes within and across cancer cell lines. We used genetic alteration data to create unique models for each cell line to improve overall performance and classify less common dependency genes that occur in a smaller subset of cell lines lines.

Based on the available project Achilles mutation and expression data we applied a number of treatments to the base PPI networks to encode each cell line’s unique genetic alteration profile as discussed below.

Mutations such as frameshift indels or nonsense substitutions were labelled as loss of function. For missense mutations the Pathogenic mutations were identified using the SIFT online (Sim et al., 2012) and then split into either loss of function or gain of function using the MoKCaRF (Baeissa, 2019) algorithm. Nodes that represented genes with inactivating mutations were removed from the PPI network, for those that represented gain of function we amended the weights of their outgoing edges as discussed below.

As well as removing inactivated nodes we weighted edges to represent the strength of the signal between the two genes – the stronger the signal the lower the barrier. Two unidirectional edges were created between each gene pair (g1, g2).

We calculated each edge weight so that as gene expression (g) tends to 0, weight (w) tends to 1. As g tends to infinity, w tends to 0. Specifically

Equation 5.2

$$w = 0.5 - 0.5 * (math.\tanh(math.\log(g + 1\text{e-}10)))$$

Where w is the weight assigned to an edge and g is the expression score for the outgoing gene node.

For genes subject to a gain of function mutation we multiplied the gene expression by 10 before calculating the weight. Whilst the exact equation w is somewhat arbitrary we found that our results were robust to changes in w.

We used three distinct versions of our expression data to implement these perturbations. We first used the raw expression data for each gene directly, next we normalised the expression level of each gene in a cell line against the same gene in all other cell lines of the same tissue type and finally, we normalised the data against the same gene in all other cell lines.

We found that of all the PPI networks treatments the raw gene expression data showed the best overall predictive performance both within and across cell lines. Within cell lines our raw data models scored a mean AUC ROC of 0.812 (s.d. 0.023) compared to the base model's performance of AUC ROC 0.765 (s.d. 0.024).

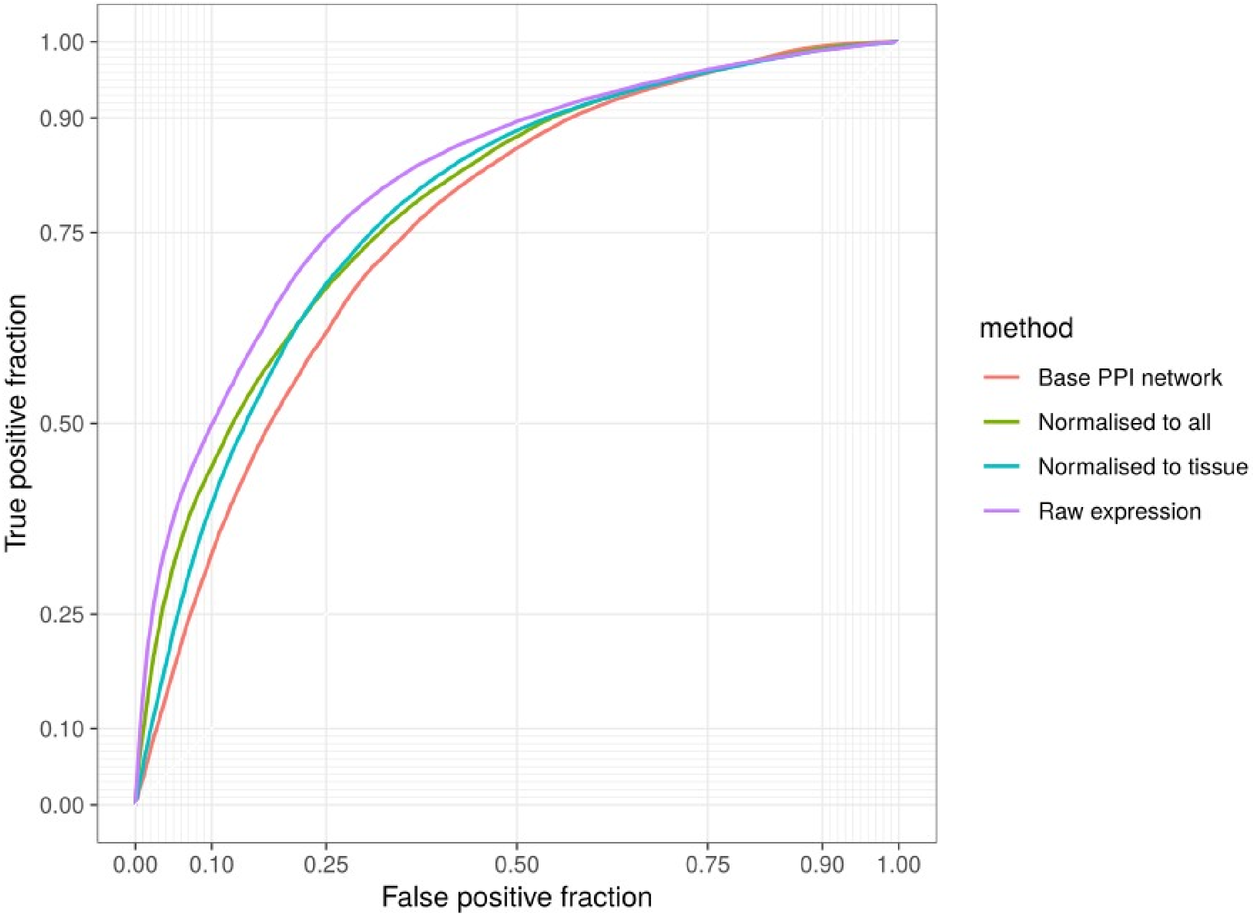


Figure 5.3. AUC ROC plots for each PPI model show that our raw expression model exhibits the largest AUC ROC, and therefor the best performance, while the base PPI model shows the worst performance.

Predicting across all cell lines and all rarities of gene our raw data model performed with ROC AUC of 0.801 (s.d. 0.006) again an improvement performance to that of the base PPI networks model's mean ROC AUC of 0.758 (s.d. 0.007) (Figure 5.3) (Table 5.3).

PPI networks treatment	Description	Within Cell line mean AUC ROC	Across all cell lines mean AUC ROC
Base PPI network	Non-directional PPI network sourced via STRING.	0.765 (s.d. 0.024)	0.758 (s.d. 0.007).
Compared to tissue type	Gene expression normalised against all cell lines of same tissue type.	0.778 (s.d. 0.021)	0.756 (s.d. 0.128)
Compared to all	Gene expression normalised against all cell lines.	0.781(s.d. 0.023)	0.756 (s.d. 0.01)
Raw data	Raw gene expression data via Cancer cell line encyclopaedia (CCLE).	0.812 (s.d. 0.023)	0.801 (s.d. 0.006)

PPI networks treatment	Description	Within Cell line mean AUC ROC	Across all cell lines mean AUC ROC
Base PPI network	Non-directional PPI network sourced via	0.765 (s.d. 0.024)	0.758 (s.d. 0.007).

	STRING.		
Compared to tissue type	Gene expression normalised against all cell lines of same tissue type.	0.778 (s.d. 0.021)	0.756 (s.d. 0.128)
Compared to all	Gene expression normalised against all cell lines.	0.781(s.d. 0.023)	0.756 (s.d. 0.01)
Raw data	Raw gene expression data via Cancer cell line encyclopaedia (CCLE).	0.812 (s.d. 0.023)	0.801 (s.d. 0.006)

Table 5.3. Mean model performance when predicting gene dependencies within each cell line (where training and test datasets were sourced from a single cell line) and across cell lines (where training was sourced from one cell line and used to classify all other cell lines). Performance measured with mean AUC ROC scores.

5.3.8 Perturbed PPI network models perform well for both common and rarer gene dependencies across cell lines

To quantify how well our models predict those genes with high dependency scores in only a few cell lines we trained our models on all cell lines and then performed validation on test sets filtered for the rarity of the acquired essential genes being predicted.

580 of the total 4030 (~14.3%) essential genes in our training data were identified as essential in all 39 cell lines. 2424 (~60.1%) were essential in more than half of the cell lines and 821 (~20.3%) of the total genes were specific to just one cell line. We created test sets featuring genes that occurred in just one cell line, below 10, 20, 30 in all 39 cell lines to calculate how well our models performed at each gene dependency rarity interval.

Of our four models, three (our base PPI network, proportional to tissue and proportional to all models) had similar levels of predictive ability for gene dependencies found in all cell lines in our training data. However across the other rarity intervals the proportional models performed slightly better than the base PPI model.

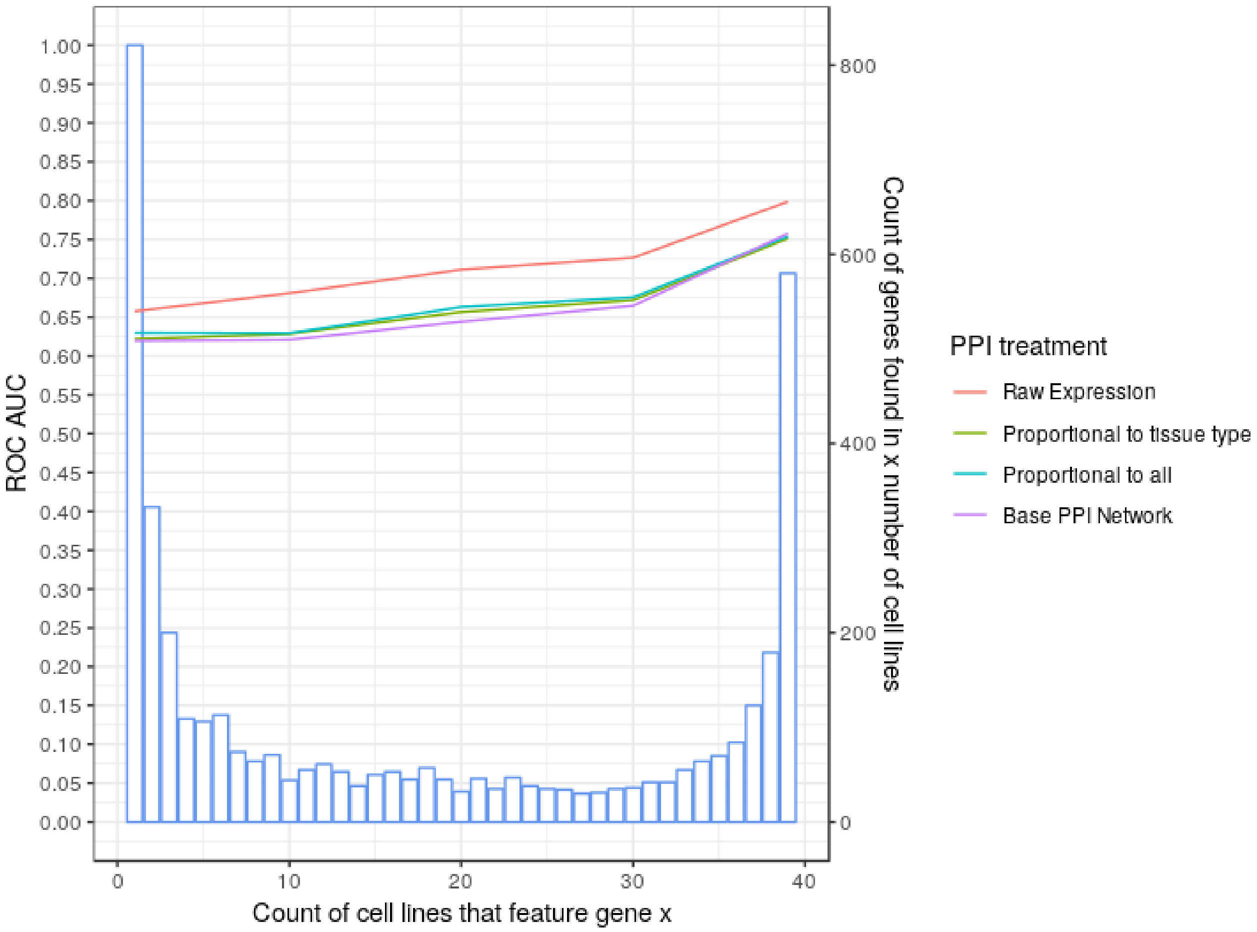


Figure 5.4. Model performance across gene dependency rarity intervals shows the general improved performance of the raw expression model. Each coloured line represent a models performance at each interval as per the legend where the blue bars representing the distribution of genes at each rarity level. For example 200 genes are reported to be dependency genes in exactly three cell lines.

The final model, our raw expression model outperformed the other models by some margin reporting a mean ROC AUC 0.660 when predicting genes that were reported as a dependency in only one cell line (compared to base model's 0.615), 0.681 for genes that showed dependency in less than 10 cell lines (compared to 0.621), 0.711

(compared to 0.644) for genes in less than 20 cell lines, 0.727 (compared to 0.665) for genes in less than 30 cell lines and 0.801 for all gene dependency rarities (compared to 0.758 for the base PPI model) (Figure 5.4.).

5.3.9 Our models are robust to PPI networks incompleteness

It is known that current PPI networks models are both incomplete and suffer from ascertainment bias in that some proteins are better studied than others (Huttlin et al., 2017; Mosca et al., 2013; Rolland et al., 2014). In order to quantify how the incomplete nature of the PPI networks affects the robustness of our models, we repeated our classification pipelines with revised PPI networks data randomly holding out 25% of the data from original network. In the case of the 25% holdout PPI networks network we observed minimal loss of predictive power from our raw expression cross cell line model with mean reported performances of AUC ROC 0.78 (s.d. 0.011) compared to 0.801 (s.d. 0.006).

We conclude that while an increasingly complete PPI network may improve our models predictive performance our current models are fairly resilient to the incomplete nature of the currently available PPI networks data.

5.3.10 Creating a pan tissue cell line training set

To maximize the amount of training data available for use by our classifiers for the prediction of gene dependencies in previously unlabelled cell lines we concatenated all available cell line training sets from all tissue types into one super set. We used our raw expression models for this super set based on their relatively high overall performance during previous validation.

In an attempt to estimate how well this concatenated data should perform for the prediction of gene dependency in unlabelled data sets we once more validated each of our individual test sets based on models trained using our super training set.

We found that our super training set classified gene dependencies across all cell lines with an AUC ROC of 0.843 (s.d. 0.012), a further improvement on the individual raw expression model's mean cross cell-line AUC ROC score of 0.801 (s.d. 0.006).

This model provided the greatest predictive power and as such represents the most suitable available for predicting gene dependencies in cell lines with no prior labelling as discussed below.

5.3.11 Predicting and validating gene dependencies in previously unlabelled cancer cell lines.

To create a set of predictions we took 37 cell lines previously unlabelled for gene dependency, 16 for breast, 13 for kidney and 8 for pancreas. Each of these cell lines was chosen based on the amount of mutation and expression training data available. We used our pan-tissue training set to train our classifiers and produced a full set of predictions for each of these cell lines.

Survival screens focusing on a library of 240 genes involved in the DNA damage response (DDR) were repeated in triplicate for the MCF7 breast cell line. Cell viability was reported using a z-score where positive numbers suggested a cell’s viability increases with the knock-down of the predicted gene, negative scores suggests a decrease in viability and z-scores below -1 constitute a true dependency. The variance of results across all three repeats was high. This may have been due to the choice of library. The loss of genes involved in the DDR can often lead to genomic instability in a cell. Knocking out a single gene (e.g. MSH3) can cause the subsequent loss of different sets of genes, resulting in different sets of dependencies.

Gene name	**Z-score**	**Dependency Likihood**
RAD23B	-0.4723	0.9741
RAD23A	0.2654	0.9713
PRPF19	-0.3052	0.9704
SHFM1	-0.3754	0.9681
TP53BP1	0.7196	0.9554
RUVBL2	-0.0575	0.9538
TRIM28	-0.6968	0.9470
XRCC5	-0.2933	0.9467
RAD1	-0.4956	0.9455
XAB2	-0.7499	0.9360

Table 5.4. Top 10 dependency gene predictions with likelihood score reported by our pan-tissue classifier and z-scores from the MCF7 DDR library survival screens. Negative z-scores suggest that the knockout of a predicted gene impacts cell viability and z-scores of below 1 suggest

dependency. 8 of these 10 genes showed negative z-score with XAB and TRIM28 reporting a z-score of less than -1 in at least one repeat of the screen.

We ranked all of our predictions for MCF7 by dependency likelihood score. Filtering for likelihood scores to keep predictions of above 0.85 and below 0.15 and treating negative z-scores as a hit we report an accuracy of 0.64 with a sensitivity of 0.73 and a false discovery rate of 0.38 based on experimental validation for the MCF7 cell line. Next, we extracted the top 10 predictions. 8 of our top 10 predictions showed signs of essentiality with a mean negative z-score. Two of these top 10 predictions, PARP1 and TRIM28, reported a z-score of less than -1 in at least one repeat (Table 5.4).

Gene name	**Z-score**	**Dependency likihood**
POLA1	-1.9200	0.6750
MEN1	-1.6886	0.8546
PNKP	-1.5129	0.5851
LIG3	-1.4379	0.3555
CHEK1	-1.2784	0.8503
EME1	-1.2168	0.4798
RBBP8	-1.2160	0.7818
PARP1	-0.9221	0.8987
ERCC2	-0.9021	0.6446
RECQL5	-0.8604	0.5324

Table 5.5. The 10 lowest genes by reported z-score in the MCF7 cell line with dependency likelihood scores given by our pan-cancer classifier. Three of these, MEN1, CHEK1 and PARP1 obtained dependency likelihoods of over 0.85 and 8 of the 10 scored over 0.5.

Only 7 of the 240 genes screened and classified for in the MCF7 cell line reported a mean z-score of less than -1 in all three repeats and two of these, MEN1 and CHEK1 were predicted as gene dependencies with a score of over 0.85 (Table 5.5).

5.3.12 Therapeutic opportunities in cancer dependency genes

Using Cansar's cancer protein annotation tools (Bulusu et al., 2014) we labelled our predicted dependency genes, based on their respective protein products, as either a drug target, druggable or non-druggable (Figure 5.5).

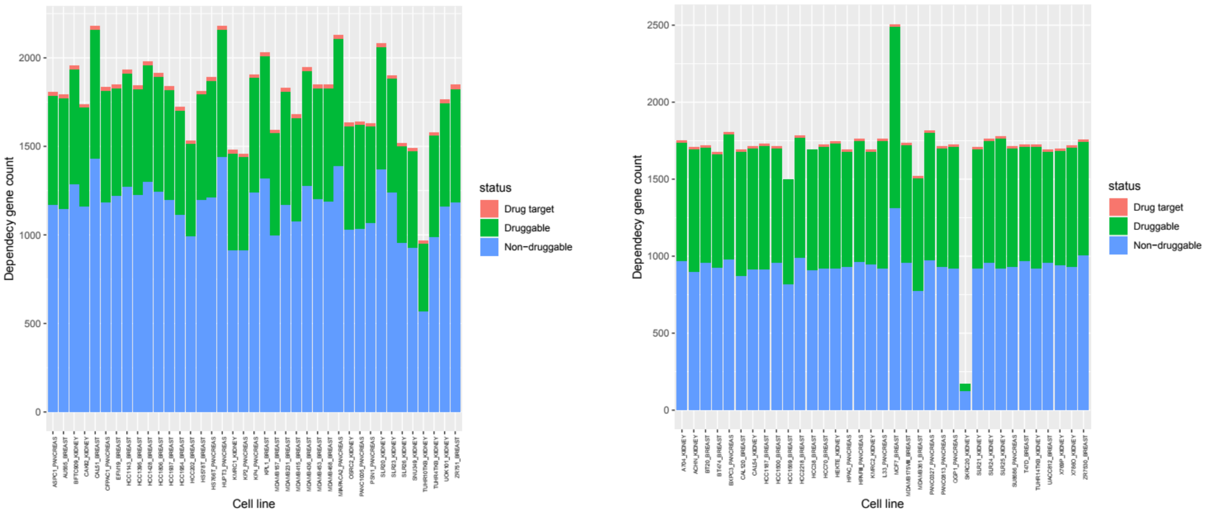


Figure 5.5. Dependency gene druggability counts by cell line. a. A histogram of dependency gene counts per cell line in our training data stratified by druggability status as reported by cansar black's cancer protein annotation tools. b. Predicted dependency gene druggability count by cell line.

The proportion of known drug targets in our predicted gene set was slightly lower than those in our training data at 0.7% compared to 1.1%. The proportion of predicted druggable genes based on a 3 dimensional structure was higher at 45.1% compared to 34.2% in our training set. We found therapeutic opportunities in almost every cell-line in both our training data and prediction set both in the form of genes with known drugs and genes that exhibit druggable traits.

5.4 Discussion

Protein-protein interaction maps provide us with a robust model of how the proteome is organised. Here we find that the topological relationships across these maps tend to be different for essential genes and non-essential genes, opening up the opportunity for predicting gene dependency. We find that topological features can be used to predict gene dependency in human cell lines with ROC AUC scores of up to 0.84. This is an improvement on accuracy reported by previous studies that use PPI network models to predict essential genes in S. cerevisiae (Saha and Heber, 2006) and E coli (da Silva et al., 2008).

Jeong et al's seminal publication (Jeong et al., 2001) was the first to show a correlation between degree centrality, I.e. the number of edges leading in or out of a given node, and gene essentiality. We find here that it is possible to use these and other topological features to predict essential genes and acquired essential genes in previously unseen cell lines, using models trained on different cell lines. We note though that the topological features that are predictive of gene dependency such as eigencentrality are predominantly measures of a protein's connectedness. These features are robust to the type of network perturbations caused by changes in gene expression and mutations. This suggests that modified PPI networks can only provide a partial picture of gene essentiality.

We described how the standard PPI Network does not capture the massive cell reorganisation seen in cancer, due to genetic mutations, copy number variances and epigenetic changes affecting gene expression. By personalising our PPI networks to reflect some of these changes we were able to model our cells lines better and improve predictive power gene dependency classification. This improvement is particularly noticeable for those genes we are particularly interested in, i.e. the genes which are essential in only a few cell lines.

Despite the relatively high performance of our classifiers we are aware that the association between gene expression and protein expression is only partial and so it is likely that further improvements will be possible for this type of model when it is possible to modify the PPI network as a result of protein expression as well as existing 'omics data.

Additionally consideration of the biological nature of the protein interactions reported as well as improvements to the completeness of our source PPI networks is also likely to lead to significant improvements in this type of study. In particular our source protein-protein interaction network provides only non-directional, binary information about interactions between proteins rather than the inhibitory or excitatory nature of the interaction. Although we report that our models are relatively robust to incompleteness in the source networks we expect that as the completeness and sophistication of PPI models improves so will the effectiveness of this type of model.

6 - Defining signatures of arm-wise copy number change and their associated drivers in kidney cancers

6.1 Introduction

One of the most striking features of the cancer cell genome is the frequently observed abnormal karyotype. Many factors can lead to abnormal structural rearrangements of chromosomes including errors in cell division such as Spindle assembly checkpoint defects (Orr et al., 2015) and missegregation due to issues such as telomerase insufficiency (Millet and Makovets, 2016). The inaccurate repair of DNA double-strand breaks can also result in translocations, duplications, deletions and inversions of DNA leading to genome instability (Aparicio et al., 2014). Recurrent translocations are frequently observed in haematological malignancies where the resulting fusion genes drive tumourigenesis (Gordon et al., 2012). Loss of heterozygosity (LOH) can also contribute to the loss of function of tumour suppressor genes (Burrell et al., 2013), and large-scale copy number changes can lead to oncogene amplification (Bagci and Kurtgöz, 2015; Schwab, 1999) .

In addition, defects in the fidelity of chromosome segregation can lead to the gain or loss of entire chromosomes. The chromosoma can vary greatly within tumours (Stephens et al., 2012) and changes in ploidy are known contributors to tumourigenesis and the progression of certain cancer types (Holland and Cleveland, 2012). Indeed some individual gene variations such as alterations in BRAF (Kamata et al., 2010), TP53 (Thompson and Compton, 2010; Tomasini et al., 2008), PTEN (Puc et al., 2005) and VHL (Thoma et al., 2009) have been linked directly with genome stability and changes in the aneuploidy state of the cell.

The production of large-scale cancer sequencing projects, such as The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015), now available via the Genomic Data Commons (GDC) (National Cancer Institute, 2019) and The International Cancer Genome Consortium (ICGC) (Zhang et al., 2011), has enabled the detailed interrogation and analysis of the cancer genomes. The multi-omic data sets generated have allowed researchers to identify driver genes (Tamborero et al., 2013), investigate correlations between expression, copy number variance (CNV) and mutation data (Gerstung et al., 2015) and to calculate the mutual exclusivity of altered gene sets (Ciriello et al., 2012).

CNVs have been used as predictive features in previous studies in an attempt to characterise distinct classes of tumour and provide insight into the functional significance of alterations across the cancer genome. Zack et al. (Zack et al., 2013) found recurrent copy number aberrations in a number of pan-cancer regions where no oncogene or tumour suppressor had previously been described. They suggested that recurrently deleted regions could either be enriched for novel tumour suppressors or enriched with non-essential genes. Kim et al. (Kim et al., 2013) revealed similarity of chromosomal arm-level alterations among developmentally related tumour types as well as a number of co-occurring pairs of arm-level alterations.

Ciriello et al. (Ciriello et al., 2013) have postulated that tumours can be classified as either M class or C class, that is, those driven primarily by mutations (M class) or copy number aberrations, often occurring alongside mutations specifically in TP53 (C class). In their study C class tumours include breast (BRCA), ovarian (OV), lung (LUSC) and head and neck squamous cell (HNSC) carcinomas. M Class tumours include kidney clear-cell carcinoma (KIRC), glioblastoma multiformae (GBM), acute myeloid leukemia (LAML), colorectal carcinoma (COAD) .

There are three subtypes of kidney cancers, renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP) and kidney chromophobe (KICH). KIRC is the most common form of kidney cancer and the 8th most common form of cancer in the UK. KIRC often presents a distinctive, previously described karyotype. KIRC tumours are often initiated by loss of function of VHL caused by a combination of mutation or epigenetic silencing of the gene in one allele, in conjunction with the loss of heterozygosity of chromosomal arm 3p, where the gene resides (Gnarra et al., 1994). For complete transformation into a cancer cell, further abnormalities are required. These commonly include mutations in PBRM1, BAP1 or SETD2 all of which are also located on chromosome arm 3p (A Ari Hakimi, Irina Ostrovnaya, Boris Reva, Nikolaus Schultz, Ying-Bei Chen, Mithat Gonen, Han Liu, Shugaku Takeda, Martin H Voss, Satish K Tickoo, Victor E Reuter, Paul Russo, Emily H. Cheng, Chris Sander and rt J. Motzer, 2013). Other characteristic arm-wise copy number changes in KIRC tumours include losses in 8p, 9p and 14q and gains in 5q and 7q (Turajlic et al., 2015).

n this study we focused on the analysis of kidney cancers, giving us the opportunity to compare and contrast three molecularly distinct cancers that all arise at the same primary site. In addition, KIRC has a distinct and well characterised karyotype, aiding the validation of our methods. We utilised publicly available copy number data to compare the overall copy number status of cancer cell lines across tissue types and to identify correlated arm-wise copy number changes and use unsupervised machine learning to discover recurring patterns of chromosomal arm ploidy change across each cell line and tissue type. Using these insights we investigated the association of commonly mutated genes in kidney cancers with the general aneuploidy status of samples across our three kidney tissue datasets. Finally, to explore whether specific mutated genes could lead to specific arm-wise copy number variance patterns, we employed a set of supervised machine learning classifiers to measure the predictive power of chromosome arm ploidy profile data to identify mutations in specific genes.

Describing patterns of copy number change across diseases and identifying associated gene mutations may provide clues for the drivers of these genomic instabilities and how specific genes interact with the karyotype as a whole.

6.2 Materials and Methods

6.2.1 Data acquisition

In total, we analysed 3,559,315 samples (356,069 in our kidney group, 3,203,246 in our pan-cancer group) across 5756 patients (888 kindey, 4868 pan-cancer). Our source data consisted of somatic mutation data (BCM Curated or Automatic Somatic Mutation Calling) and copy number variant data (BI Genome-Wide SNP6) downloaded from the GDC data portal (National Cancer Institute, 2019).

For our kidney cancer data set we included samples from KIRC (Renal Clear Cell Carcinoma), KIRP (Kidney renal papillary cell carcinoma) and KICH (Kidney Chromophobe). For our pan-cancer dataset we included patients from cohorts with tissue types including; BRCA (Breast cancer), DLBC (Diffuse large B-cell lymphoma), GBM (Glioblastoma), LGG (Low Grade Glioma), LUSC (Lung Squamous Cell Carcinoma), MESO (Mesothelioma), OV (Ovarian), PRAD (Prostate Adenocarcinoma), SARC (Sarcoma), UCEC (Uterine Corpus Endometrial Carcinoma).

To compare the aneuploidy exhibited by different tumour types we used the segment mean values in the CNV data provided by TCGA / GDC. The segment mean value is the log2 geometric mean of the ratio a sample's copy number over the wild-type copy number. Throughout this study, our CCN values were calculated directly from these segment mean values using the formula:

Equation 6.1

$$\left(2^{sm}\right)*2$$

Where sm is segment mean.

Sex chromosomes were excluded from the study, thus the chromosomal copy number for healthy tissue is 2.

To ensure the data represented large-scale changes in ploidy rather than smaller, linear duplication CNVs, we sorted the sample data by length and removed the shortest samples observed in the bottom 25th percentile of data. Samples with a probe number of less than 10 were also filtered out to provide high confidence in the CNV data as per Laddha et al. (Laddha et al., 2014).

All data preprocessing and analysis was completed using R 3.4.4 (R Development Core Team, 2011) and the machine learning component of the study was completed using python pandas version 0.19 (McKinney, 2010).

6.2.2 Describing chromosome arm-wise ploidy patterns

To prepare the data for modelling, we segmented CNV data for each patient by chromosome arm. Centromere position data were obtained from genome.ucsc.edu (James Kent et al., 2002). These data were organised with each patient as an observation and each of their chromosome arm's average CCN as a variable.

6.2.3 Correlation of mutated genes with overall ploidy change

We identified the ten most recurrently mutated genes in KIRC and KIRP tumours via cBioPortal (Cerami et al., 2012). In KIRC tissue these genes included VHL (occurring in 47.8% of KIRC patients), PBRM1 (34.8%), SETD2 (12.1%), BAP1 (~9.4%), MTOR (6.7%), KDM5C (5.9%), ARID1A (4.1%), KMT2C (4.1%), SPEN (3.8) and PTEN (3.8). In KIRP these genes include MET (7.4%), KMT2C (6.4%), SETD2 (5.7%), KMT2D (5.0%), BAP1 (5.0%), AR (4.6%), FAT1 (4.3%), PCLO (4.3%), PBRM1 (3.9%), NF2 (3.5%). Little mutation data was available for KICH at the time of writing.

Samples with deleterious mutations that were likely to change the protein product or the functioning of the protein such as missense mutations, frameshift insertion and deletions, nonsense mutations and in-frame insertions and deletions were labelled as mutations in the data whilst all other samples were labelled as not-mutated.

To calculate correlation between each mutation and overall genome ploidy we took an absolute, positive value for each sample's segment mean to better measure and compare overall average ploidy change. This data was grouped by individual patients resulting in a list of all patients each with an average absolute genome-wide segment mean.

A list of patients featuring mutations in our chosen genes was then extracted from the somatic mutation data and used to filter our preprocessed CNV data to calculate the probability of each chosen gene mutation being associated with a general change in ploidy. For each gene, the cohort of patients with a mutation in that particular gene was sampled against a control group of all other available patients.

We used t-tests to calculate the probability that patients with mutations in each gene would exhibit a significant change in overall ploidy. This was repeated in our pan-cancer, KIRC and KIRP tissue datasets for comparison. All p-values were adjusted through the use of the Bonferroni procedure to correct for false discovery.

6.2.4 Arm ploidy correlations

To measure correlation between chromosome arms we first stratified patients by tissue type and then measured correlational coefficients between each arm pair for all the samples within those groups.

Significance values were calculated using a permutation test where a mock distribution was produced using the same data for each arm pair but with one arm's data randomly permeated for 1000 samples. The real value was then compared to this mock distribution using a student t-test.

6.2.5 Generating arm-wise ploidy signatures

Arm-wise ploidy signatures were generated from the arm-wise ploidy data (as above) using non-negative matrix factorisation (NMF). We used a Cophenetic Correlation Coefficient score via the R NMF library (Gaujoux and Seoighe, 2010) to measure the stability of our models and to select the most stable component count. When associating a sample with a signature we chose the highest scoring component in the coefficient matrix for that sample.

6.2.6 Gene mutation and pattern change

To measure the distance between the arm-ploidy pattern of groups of patients with and without gene mutations we took the median values of each arm for each group and measured distance using cosine similarity. We took the top 250 most frequently mutated genes in kidney cancers to analyse. Similar to the arm ploidy correlation analysis, we used permutation tests to measure significance for each of these gene mutations, this time randomising the labelling of patients in the two groups.

6.2.7 Finding gene mutations enriched within signatures

To find the most frequently mutated genes in each signature we stratified our samples by most prominent signature (as above), counted the frequency of gene mutations in these groups and ranked by their frequency. To determine whether the frequency of mutations was significantly increased compared to that which we would expect at random we once again used permutation tests using randomised sampling of all patients from all signatures to create our mock distribution. In each case the number of randomly sampled patients was equal to the number of samples found in the respective signature.

6.2.8 Using chromosome arm ploidy patterns to predict gene mutations

A boolean, stating whether the patient suffered a mutation in the respective gene or not, served as the label for each observation. Data was split randomly into training and testing groups with a test size of 0.2 and a training set size of 0.8 of all samples.

To measure the predictive power of chromosome arm-wise segment mean for specific gene mutations, receiver operating characteristic area under the curve (ROC AUC) scores were calculated. We initially trialled four different machine learning classifiers;

Bernoulli Naive Bayes, Support Vector Machine, logistic regression and random forest. Hyper-parameters for each classifier were optimised using 5 fold cross validation.

Ultimately it was found that a random forest classifier with 1000 estimators and a minimum sample leaf size of 30 performed consistently better when compared to the other classifiers and so ROC AUC scores given in this study were all a result of this classifier

For each model, feature importance was calculated by measuring the mean decrease in classifier accuracy with the removal of each feature across all trees in the random forest. This metric was reported as the mean decrease in accuracy given the removal of a feature.

Using this measure of the mean decrease in accuracy we ranked and identified which chromosome arms were most commonly lost or gained when specific genes are mutated. This analysis was applied to pan-cancer data, KIRC tissue data alone and finally to all tissue types excluding KIRC for comparison.

The source code and data for this study is available via bitbucket at https://bitbucket.org/bioinformatics_lab_sussex/ploidy_nmf. Results data is available at https://bitbucket.org/bioinformatics_lab_sussex/ploidy_nmf/downloads/.

6.3 Results

6.3.1 Magnitude of copy number changes differs between cancer subtypes

We first compared overall chromosomal copy number change across all tissue types. We processed raw CNV data downloaded from the GDC data portal (National Cancer Institute, 2019). This data was originally generated by using Affymetrix Genome-Wide Human SNP Array 6.0 to identify repeated regions of the sample genome and to further infer the copy number of these repeats. We calculated mean chromosomal copy number (CCN) for all samples (excluding sex chromosomes) and grouped these samples by cancer subtype (Figure 6.1).

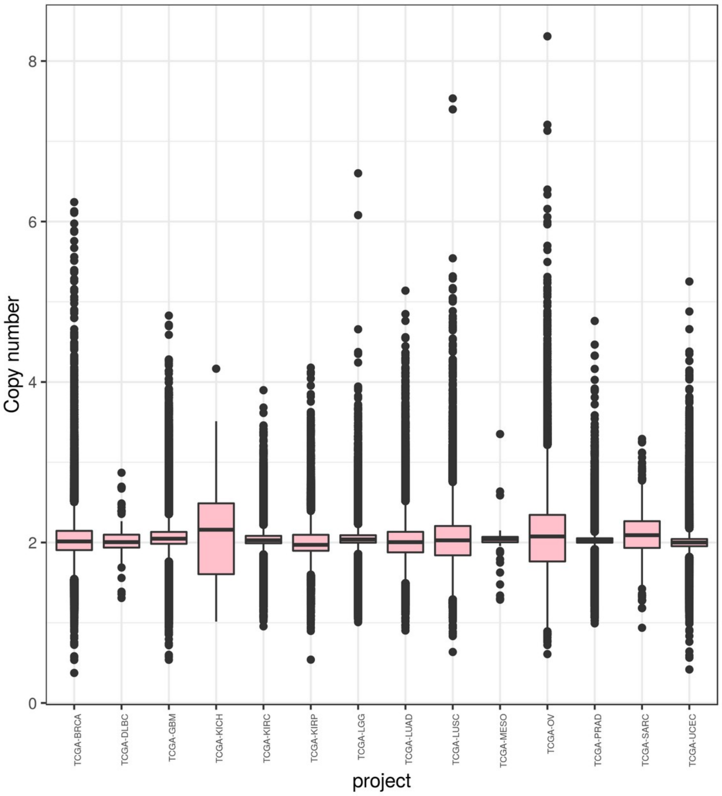


Figure 6.1. Chromosomal copy number by tissue type – A box-plot summarising mean genome-wide copy number for all samples in the 14 tissue types featured in our pan-cancer dataset. Copy number values were converted from segment mean sourced via TCGA / GDC data.

Across tumours from all tissue types present in this study, we see a slight overall gain of genetic material with a mean CCN of 2.046. There is much variance with a bottom quartile of CCN 1.932 and a top quartile of CCN 2.128. The lowest CCN recorded in our samples is 0.375, from a breast cancer sample, and the highest 8.308 from an ovarian cancer sample.

Mean chromosome copy number did not fall below 2 for any of the tissue types included in this study. KICH exhibited most gain with a mean CCN of 2.105 (s.d. 0.450) as well as the lowest 25th percentile at CCN 1.605 and the highest 75th percentile at CCN 2.488.

Ovarian cancer samples showed the most variance of CCN with a standard deviation of 0.499. The tissue types with the lowest variance of CCN included prostate adenocarcinoma and KIRC with standard deviations of 0.175 and 0.219 respectively.

These results show that the extent of genomic ploidy change seems to vary by cancer subtype type rather than by the primary site of the sample. For example the three classes of kidney tissue exhibited notably different distributions.

6.3.2 Different genes are associated with changes in overall aneuploidy in different tissues

We next investigated how gene mutations were associated with overall ploidy change. We identified the 250 most commonly mutated genes in our kidney cancers. Genetic mutations in VHL, PBRM1, SETD2, BAP1, MTOR, KDM5C, PCLO, KMT2C, ARID1A and SPEN were most commonly seen in KIRC and MET, KMT2C, AR, FAT1, PCLO and NF mutations were most commonly found in KIRP tissue (Cerami et al., 2012). Mutational data was not available for KICH tissue.

Within each tissue type we compared the distributions of copy number change between samples with a specific mutation and those without a mutation to calculate a significance score for each gene. After applying Bonferroni correction to correct for false discovery error we ranked these genes by the reported corrected significance score.

In KIRC tissue samples those with POLE mutations were reported to show the most significant loss in ploidy ($p = 1.53*10^{-13}$) while samples with mutations in TP53 exhibited the most significant gains ($p\ 1.33*10^{-12}$). Other prominent gene mutations associated with ploidy change in KIRC included GRM8 (associated with loss, $p = 3.44*10^{-12}$) and

SYNE1 (associated with gain, $p = 2.36*10^{-11}$) and ASTN1 (associated with gain, $p = 2.66*10^{-11}$) all of which are associated with brain function and have not previously been associated with ploidy change or KIRC tissue. Mutations in VHL and SETD2, genes traditionally associated with KIRC tissue, also result in a significant change in ploidy in our KIRC tissue samples ($p = 0.002$ and $p= 0.006$ respectively) though mutations in PBRM1 did not.

In KIRP tissue we found that patients with KRAS mutations showed the most significant loss in copy number ($p < 2.2*10^{-16}$) followed by THSD7B ($p = 2.48*10^{-36}$) and CHD4 ($p = 3.31*10^{-24}$). Patients with mutations in EP400 ($p = 8.99*10^{-16}$) and PCDH11X ($p = 2.79*10^{-09}$) exhibited the most significant levels of ploidy gain. Surprisingly we found that mutations in TP53 did not appear to result in any significant ploidy change in the available KIRP tissue samples.

6.3.3 Chromosome arm ploidy patterns vary by tissue

We next investigated copy number data stratified by chromosome arm to ascertain whether certain chromosome arms were preferentially gained or lost in the kidney cell lines compared to other cancers (Figure 6.2). All p-values below were calculated using segment mean values to allow direct comparison.

Across our kidney tissue samples we found that each tissue type exhibited a distinct pattern of copy number variance. This suggests that arm-wise copy number change profiles depend on tissue type more than the primary cancer site.

As expected based on previous studies 5q showed the highest ploidy gain on average in our KIRC tissue samples (CCN=2.27, STD=0.24) and 3p the most loss by a significant margin (CCN=1.61, STD=0.23). In comparison, in our pan-cancer data neither 5q

(CCN=1.98, STD=0.26 P=3.50*10^{-144}) or 3p (CCN=1.94, STD=0.26, P=2.21*10^{-144}) show any notable copy number variance on average.

In KIRP tissue 7p and 7q both exhibit a significant gain (both CCN 2.48, STD 0.46) whilst in our pan-tissue data the arms show less dramatic change (7q CCN=2.24, std = 0.35, 7p CCN = 2.23, std = 0.39).

KIRP also exhibited loss in 22q (CNN 1.84, STD 0.22) while KICH tissue showed a gain (2.33, 0.317, P = 1.43745*10^{-20}) compared to the pan cancer (1.92, 0.29, P = 1.71014*10^{-38}).

The KICH arm-wise copy number data exhibit much larger copy number changes across every arm with only 21p appearing relatively diploid. On average each of the other arms seems to have either gained or lost roughly half of its genetic material.

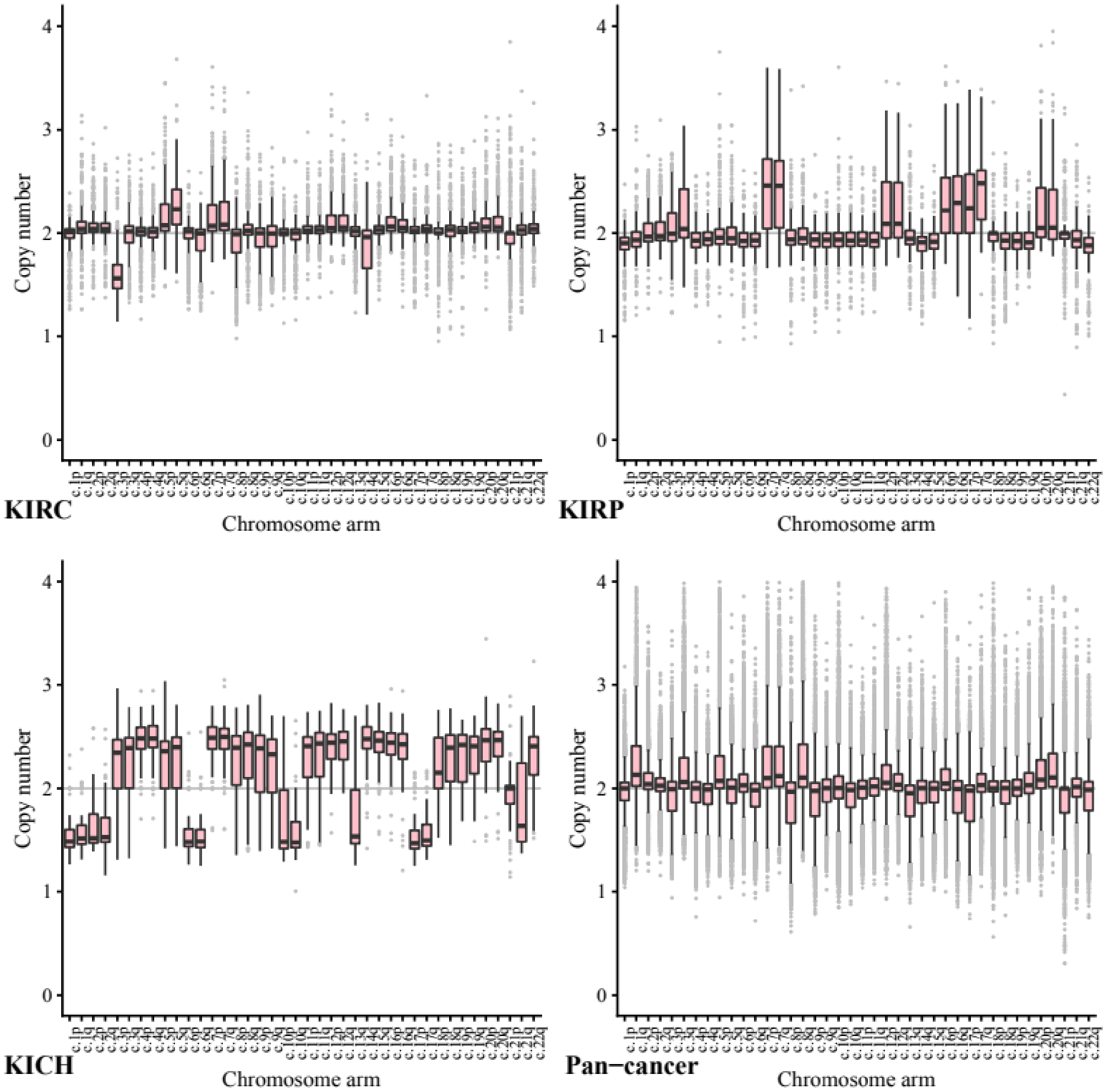


Figure 6.2. Chromosome Arm-wise copy number - By grouping chromosome arm-wise copy number data across our kidney tissues individually we develop a clearer picture of the pattern of chromosome arm copy number that occurs specifically in each kidney tissue compared to our pan-cancer dataset.

These results support existing research into the karyotype of KIRC tissue, including loss in 3p and gain in 5q.

Across the pan-cancer samples, a set of 10 cancer types excluding the three kidney cancers (see methods) the chromosomes that exhibited the most notable copy number change included 8q and 17p. We note that the patterns found in our pan-cancer samples were not shared with our kidney tissue samples.

Chromosome arm 8q was to shown to exhibit the highest average copy number gain and variance in our pan-cancer data (CCN 2.29, STD 0.0.49). Comparatively KIRC and KIRP tissues both showed significantly less change in arm 8q with a mean copy number close to normal (KIRC CCN 2.04, STD 0.22, P=1.65*10^{-87} and KIRP CCN 2.00, STD 0.25, P=2.12*10^{-55}).

Across our pan-cancer samples chromosome arm 17p showed the greatest loss of ploidy on average (CCN=1.88, STD=0.30). Again our KIRC tissue samples exhibited as near diploid (KIRC CCN=2.01 STD=0.16, P=0.17*10^{-74}) and KIRP exhibits a significant gain in copy number (KIRP CCN=2.27, STD=0.37, P=7.79*10^{-55}).

The apparent stability of 8q in KIRC and KIRP tissues compared to our post-cancer set, may be a possible avenue for further research.

6.3.4 Patterns of arm ploidy correlation are found within tissue types

A loss in chromosome arm 3p paired with a gain in chromosome arm 5q is a common and well described trait of KIRC tissue. Next we investigated how changes of ploidy in other pairs of chromosome arms relate to each other. We measured the correlation coefficient for each pair of arms within tissue type (Figure 6.3). Significance scores were measured using permutation tests as discussed in methods.

The most prominent positive correlation scores occurred between arms of the same chromosome (For example chromosome arms 1p and 1q commonly exhibited a similar amount of gain or loss together), however these scores were not uniform across all tissue types. In KIRC, for example, the 9p-9q arm pair was reported to show the highest positive correlation coefficient (r = ~0.73, p = 4.08*10^{-72}) while in KIRP it was the 20p-20q arm pair (r = ~0.78, p = 7.27*10^{-53}). Conversely the 9p-9q arm pair in KIRP exhibits a much lower correlation score (r = ~0.313, p = 2.69*10^{-18}) and the 20p-20q arm pair in KIRC shows a slightly negative correlation (r = -0.02, p = 4.17*10^{-55}).

In terms of negative correlation, where one arm's gain of genetic material was commonly associated with a loss in the others, we found the 3p-5q pair featured prominently for KIRP tissue (r = -0.6348595, p = 1.369225*10^{-07}) but perhaps surprisingly not as prominently for KIRC (r = -0.1250279, p = 2.212398*10^{-05}) although both are reported to be significant. In KIRC tissue the 4q-6q exhibits the strongest negative correlation ploidy (r = -0.5838217, p = 3.641906*10^{-05}) (Supplementary Figures 6.1 & 6.2).

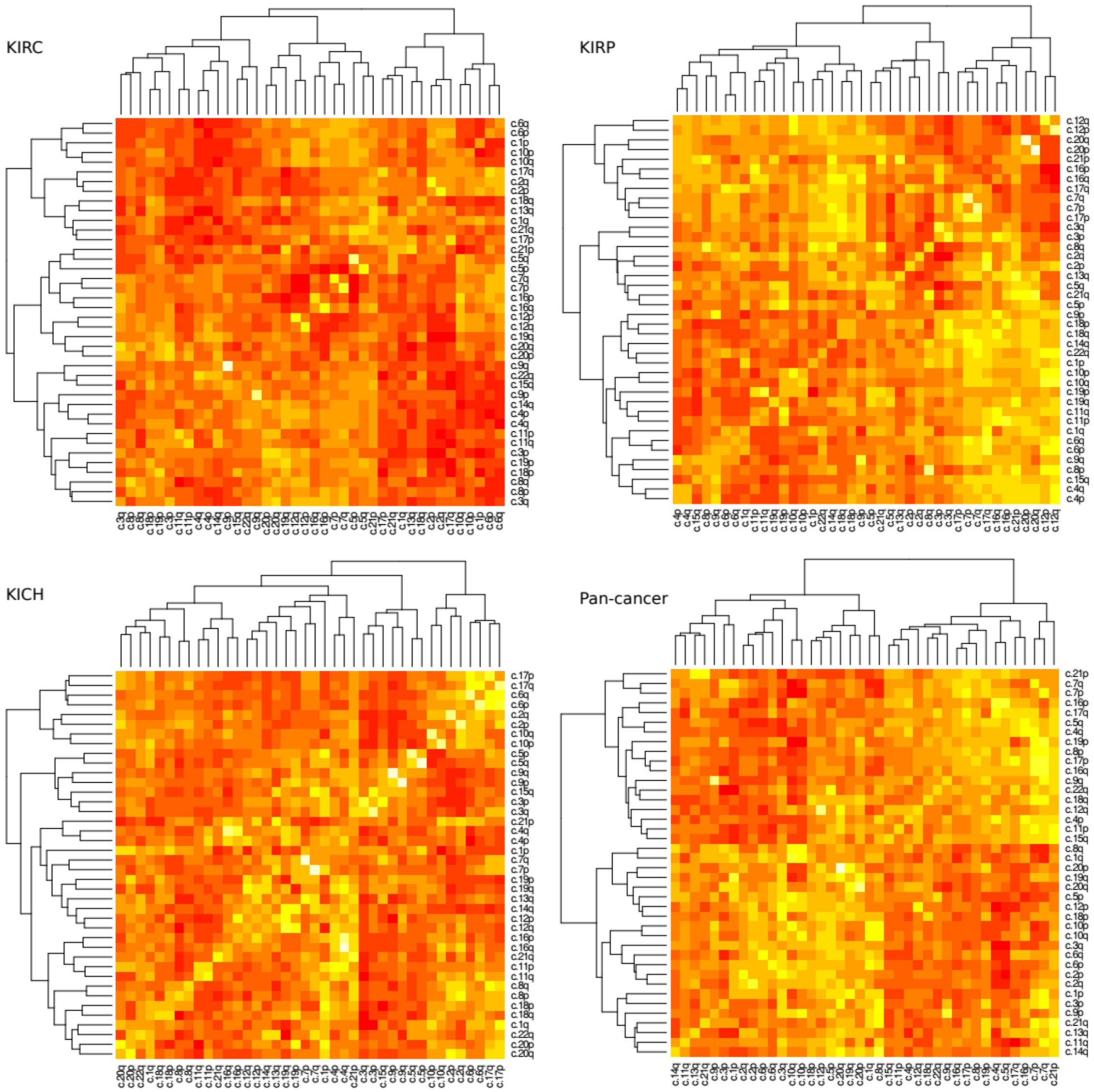


Figure 6.3. Arm ploidy correlational heat maps – Taking the Pearson correlation coefficient as our measure we visualise the correlation of gain and loss of copy number for each chromosome arm pair in each of our kidney datasets and in our pan-cancer data. Bright yellow tiles denote highly positive correlation and dark red tiles denote highly negative correlations. The results along the diagonal (which report, for example, of the correlation of c.1p – c.1p) are normalised to r=0 to improve overall contrast)

6.3.5 Arm ploidy signatures are somewhat enriched for tissue types

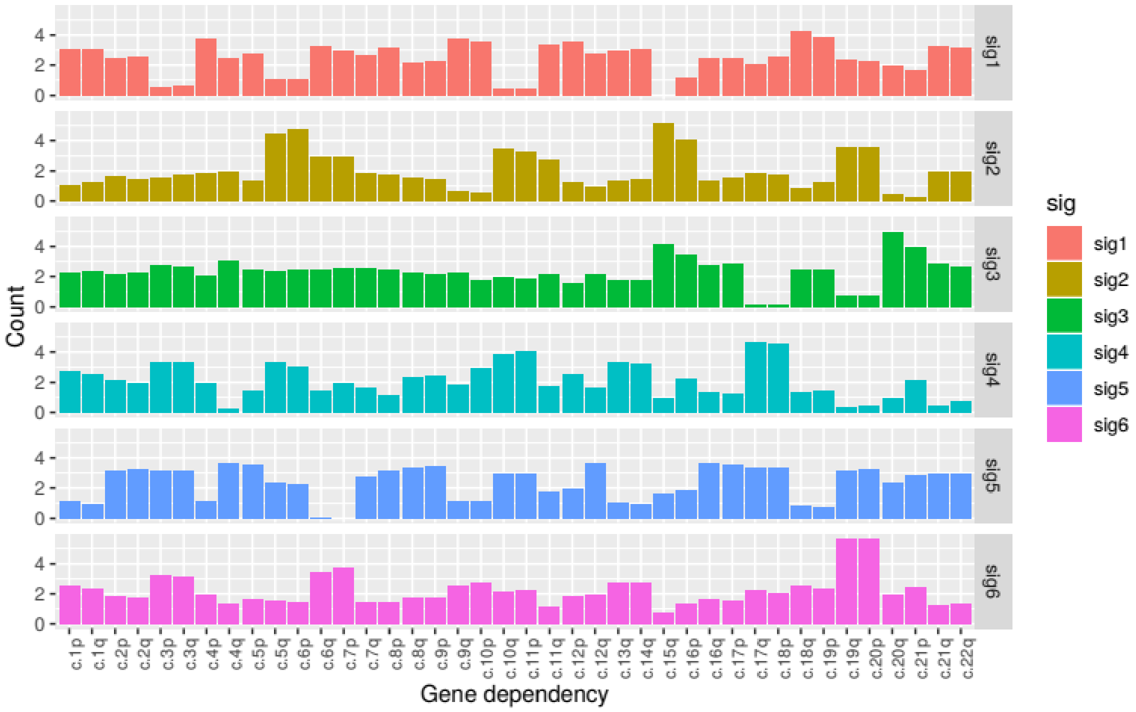
The patterns of arm ploidy correlation found above suggest that underlying mechanisms that occur in specific tissues may give rise to different profiles of loss and gain of ploidy in different chromosome arms. In an attempt to capture these profiles more fully we employed non-negative matrix factorisation (NMF) to generate a number of arm ploidy signatures across our kidney tissue data.

NMF is a multivariate analysis tool commonly used for easily interpretable decomposition. NMF was chosen because it provides both dimensionality reduction and clustering. Essentially NMF decomposes a feature matrix into two descriptive matrices, a basis, which describes the feature composition of each component and a coefficient, which describes the component composition of each sample in the original feature matrix.

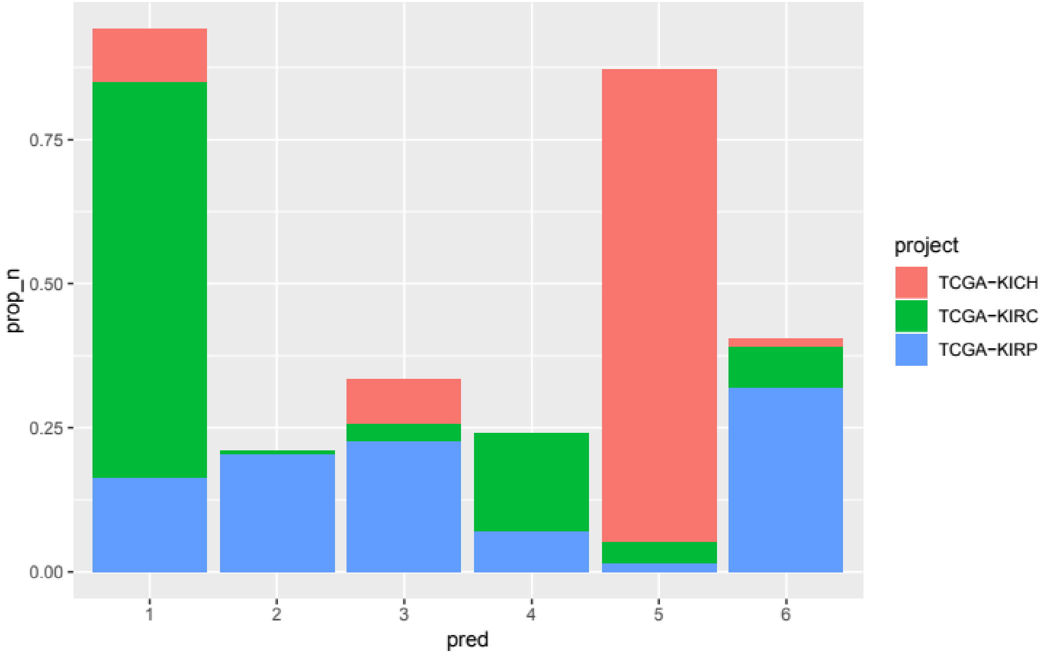
NMF generates a specified number of components (used as signatures). To identify the optimum number of components for each dataset we ran a number of trials for decompositions with an increasing value of required components. Each of these trials provided a cophenetic score, a measure cluster stability, i.e. how well the clusters obtained by NMF preserved the pairwise distances between the original data points. For each dataset we selected the lowest component count associated with a local minimum cophenentic score.

Our kidney tissue ploidy data was decomposed into six component signatures (Figure 6.4). We found that some of these signatures where highly enriched for just one of the three tissue types within the kidney group, KIRC, KIRP and KICH whilst others were more mixed. KIRC samples were most prominent in signature 1 and 4, KIRP samples were dominant in signatures 2, 3 and 6, and KICH was found most prominently in signature 5.

In terms of composition we found that the feature composition of signature 2, associated with KIRP tissue, exhibited the most notable deficiency of 3p paired with gains in 5q which reflected our correlation analysis. This was paired with large gains in 6p, 15q and 16p.



a.



b.

Figure 6.4 a. Copy number signature composition - A breakdown of the features that describe each of our 6 signatures. For example we can see that chromosome arms 19q and 20p are prominent features in signature 6. b. A chart showing the proportion of kidney patients categorised by most prominent signature and stratified by tissue type. From this visualisation we can see that signature 1 is highly enriched for KIRP tissue and signature 5 is highly enriched for KICH tissue.

We found that signature 1, most prominently associated with KIRC, featured gains in a number of arms most notably in 10p,18q and 19p. Signature 5, enriched for KICH samples, also features gains in a number of chromosome arms that do not feature in the other signatures and notably exhibits losses in arms 6p, 6q and 7p. Notably, signature 6, the other signature associated with KIRP tissue, exhibits gains in 3p and 3q paired with some loss in 5p and 5q. 19q and 20p also exhibited large gains in signature 6.

6.3.6 The amount of signature composition change varies by gene mutation

We next used cosine similarity, the cosine of the angle between two signatures represented as vectors projected in a multi-dimensional space, to measure the effect of specific mutations on overall signature composition change. By taking the median signature composition of patients with and without specific gene mutations and measuring the cosine similarities of these compositions we were able to rank genes by the distance of signature composition between these groups.

In our kidney cancer data we found a number of gene mutations associated with an increased distance in signature composition. The genes with the highest cosine similarity (all with cosine similarity scores of ~0.678) in our ranked list are included in table 6.1a.

In our pan-cancer data we found a different set of genes with some similar functionality listed in 6.1b.

We note the frequency of gene associated with the innate immune system as well as KDM6B, INCENP and H2AFJ which are all related to the chromosome organisation.

6.3.7 Some gene mutations are enriched in our kidney ploidy signatures

To find gene mutation enrichment in our six kidney cancer signatures we first categorised samples into signature groups and counted the frequency of mutations that occurred in patients in each group. We then used permutation tests to measure if each mutation in each group was significantly higher than would be expected in a random sample of patients. Below we highlight some of the notable gene mutations that were reported to be significantly enriched in each signature.

Signature 1, the largest signature group, and associated with KIRC tissue, exhibited significant enrichment for MUC4 (p = 0.008800434). MUC4 is associated with changes in ErbB2 expression, apoptosis, proliferation, differentiation, and some cancers.

In signature 2, which is enriched for KIRP tissue and composed of notable ploidy change in 3p, 5q, 6p, 15q and 16p, CENPF (Centromere protein F) was mutated in 4.2% of patients (p = 9.30e-05) and DSPP (Dentin Sialophosphoprotein), a gene associated with calcium ion binding and extracellular matrix structural constituent, was also mutated in 4.2% (p = 3.55e-05).

Signature 4, a group again associated with KIRC and changes in 3p, 3q, 17q and 18p, saw enrichment for KMT2D (also MLL2), a histone methyltransferase, which was mutated in 6.2% of patients (p=1.678634e-08).

Signature 6, associated with KIRP and changes in 6q, 7p, 19q and 20p, was enriched for MUC4 (Mucin 4) which occurred in 7.1% of patients (p = 4.890953e-06). Signatures 3 did not exhibit any significant enrichment for mutations.

Gene	Functional associations
PON2	Oxidative stress protection Pathogenic bacteriaprotection
IL6ST	Adipogenesis Innate immune system
KDM6B	Chromatin organisation control Gene silencing
SEC16B	Organisation of transitional endoplasmic reticulum sites and protein export
INCENP (Inner Centromere Protein)	A component of the chromosomal passenger complex (CPC), a key regulator of mitosis

Table 6.1a. Gene mutations associated with an increased distance in signature composition between wild type and mutated kidney cancer samples

Gene	Functional associations
RNF185 (Ring Finger Protein	Ligase activity

185)	selective mitochondrial autophagy
TMEM30C	Innate immune system
CUEDC2	DNA damage response
H2AFJ	Histone H2A.J
PRG3	Innate immune system

Table 6.1b. Gene mutations associated with an increased distance in signature composition between wild type and mutated pan-cancer samples

6.3.8 Patterns of ploidy of chromosome arms are associated with specific mutated genes

To investigate to what extent mutations in specific genes lead to specific patterns of arm-wise copy number change we trialled a set of four machine learning classifiers designed to predict gene mutations based on a patient's chromosome arm ploidy profile (see methods). Using an AUC ROC score of 0.70 as a cut-off we found that the patterns of ploidy in chromosome arms had varied predictive power for mutations in specific genes.

We trialled four classifiers; Bernoulli naive Bayes, support vector machine, logistic regression and random forest. Due to consistently better performance when compared to the other classifiers all ROC AUC scores below are based on the results of the random forest classifier. In our pan-cancer dataset patterns of ploidy changes in chromosome arms were strongly predictive of mutations in VHL (ROC 0.91), TP53 (ROC 0.74), and PBRM1 (ROC 0.71), with BAP1 (ROC 0.67) narrowly missing the cut-off score.

To better understand how different features contributed to the predictive power of each classifier we ranked the importance of each feature for each model. Feature importance was calculated by measuring the mean decrease in classifier accuracy when systematically holding out each variable across all tree permutations in a random forest. Feature importance was reported as a mean decrease in accuracy.

When analysing the important features in the predictive model, losses in 3p featured heavily in patients with BAP1, PBRM1 and VHL mutations. Gains in 5p were an important feature for both PBRM1 and VHL mutated genes.

In general KIRC tissue arm-wise CCN data, which exhibited less variance than KIRP tissue, were less predictive of specific gene mutations. The highest scoring genes, in terms of ROC AUC score, were PTEN (ROC 0.68), PBRM1 (ROC 0.61) and AKAP9 (ROC 0.61) (Figure 6.5). 18q previously noted for its overall low ploidy change in KIRC tissue was found to be the most important feature for both AKAP9 (0.15 mean decrease in accuracy) and PTEN (0.081 mean decrease in accuracy).

Chromosome arm 3p, the arm on which both PBRM1 and BAP1 are located, commonly exhibits loss in KIRC tissues. However we found that 3p reported a relatively low overall importance score in our KIRC tissue predictive models. Instead of 3p the highest ranking feature for PBRM1 was 10p (0.063 mean decrease in accuracy). This may be due to the all but uniform loss of these genes along with 3p in KIRC tissue (94% of patients with overall loss in 3p and 71% with less than CCN 1.8) leading to less variation and, as such, less signal.

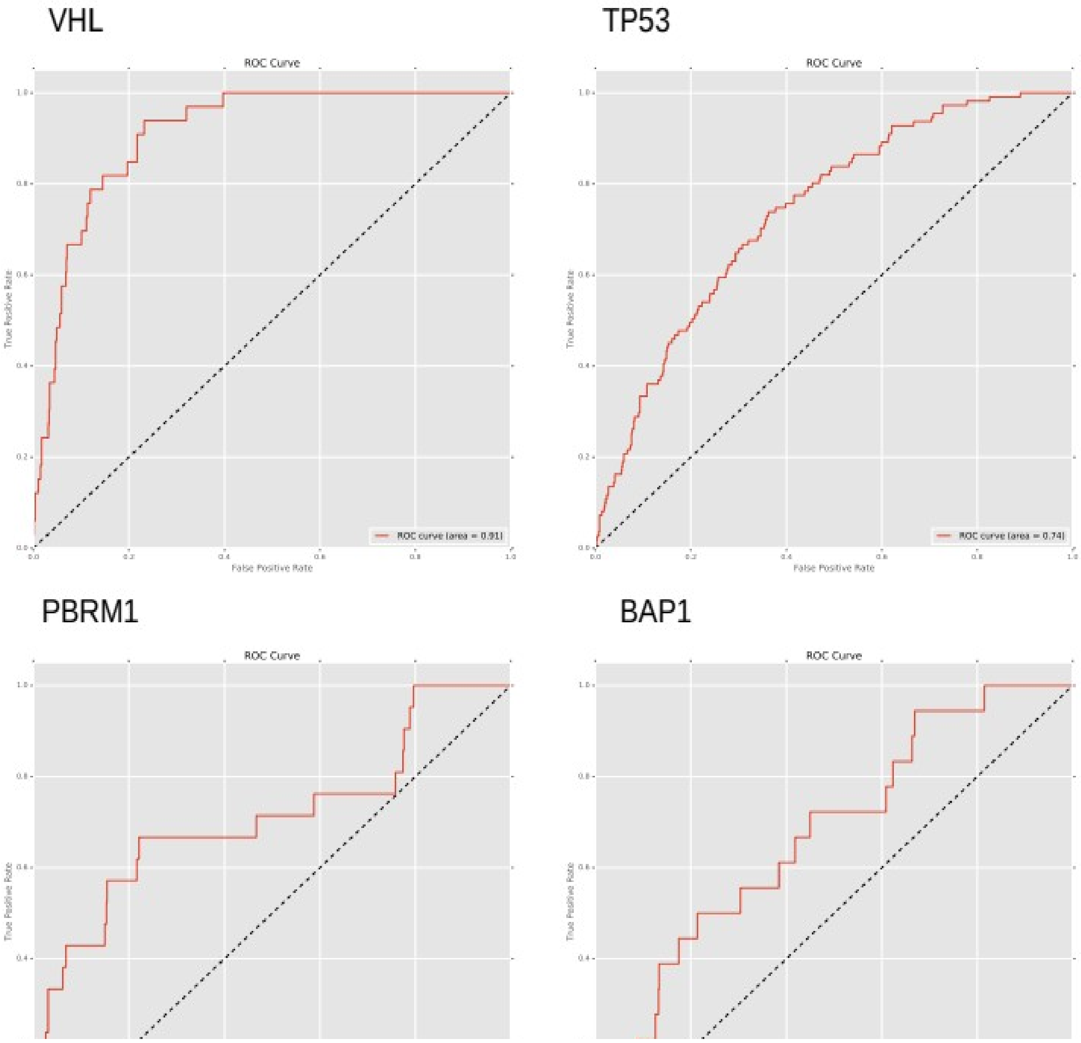


Figure 6.5. AUC ROC curves used to measure the performance of random forest classifiers trained on arm-wise chromosome copy number patterns to predict gene mutation status of VHL, TP53, PBRM1 and BAP1. A larger area under the ROC indicates better performance.

In models derived from our pan-cancer data, features based on patterns of arm-wise CCN performed relatively well when predicting TP53 and PTEN mutations with AUC ROC scores of 0.74 and 0.73 respectively. TP53 has been implicated in genetic instability in many tissue types (Donehower et al., 2019; Eyfjörd et al., 1995). BAP1 also performed fairly well with a ROC AUC score of 0.66; with the important features being a loss of 3p (0.076 mean decrease in accuracy) and of 3q (0.15 mean decrease in accuracy). Our pan-cancer tissues models performed poorly when predicting PBRM1 with a ROC AUC score of 0.47.

Once again both PBRM1 and VHL were found to be associated with KIRC tissue with losses in 3p and gains in 5p featuring as important predictors for both PBRM1 and VHL mutated genes. Chromosome arm ploidy proved a strong predictor of mutations in VHL, TP53 and PBRM1 in our pan-cancer dataset and the insight drawn from our feature importance scores again matched our expected results.

6.4 Discussion

The goal of this study was to investigate how genome-wide and chromosome arm-wise ploidy varies by tumour type, how these ploidy patterns are associated with genetic mutations and how suitable ploidy data is as a predictor of specific mutations.

As described above we observe significant variation in both genome-wide and chromosome arm-wise ploidy between samples from different tissue types. This may be expected given that different tumours are driven via gains or losses of specific genes located on various chromosome arms. Whilst we observed a small number of generalisations regarding the pattern of genetic material e.g. 20p shows gains in CCN across all tumour types, beyond these similarities, there is significant variation in arm ploidy profiles between tumour types.

Our focus on KIRC and kidney tissue was due to it is distinct and well-described karyotype. Throughout this study, we compare our observations with previous experimental studies of KIRC to cross-validate our analysis. Turajlic et al. (Turajlic et al., 2015) found that KIRC is generally characterised by recurrent copy-number variants in arms including, but not exclusive to, 3p, 5q, 7q, 8p, 9p, and 14q. SQSTM1, a gene which resides on 5q, an arm which shows relatively large gains in CCN in this study has been postulated to represent an alternative mechanism for activation of mTORC1 (Li et al., 2013). PBRM1, VHL, BAP1 and SETD2 all reside on 3p, a copy of which is known to be often lost at the outset of KIRC tumourigenesis. As such we expected changes in 3p ploidy to feature prominently in our results. The association between 3p and KIRC was supported by the findings in this study, both whilst investigating the chromosome arm ploidy profile of KIRC compared to other tissues and whilst calculating feature importance as part of our classification where we clearly found 3p performing as an important indicator for both PBRM1 and VHL mutations. We also found some association between 3p and 5q in KIRP tissue samples both in our correlation and signature analysis.

Of further interest is the prominence of 5q in our results which often sees a gain in ploidy. This gain might be a direct or indirect result of the loss of heterozygosity and resulting vulnerability to loss of gene function at 3p (Turajlic et al., 2015). Our random forest importance scores suggest that 5p is highly associated with both PBRM1 (0.088 mean decrease in accuracy) and VHL (0.13 mean decrease in accuracy) mutations across all tissues. Chen et al. (Chen et al., 2016) describe the gain of 5q as a common trait of papillary-enriched KIRC subtype which often includes mutation or amplification of MET.

VHL, a gene often found mutated early during the progression of KIRC and found associated with changes in ploidy in this study, has been shown to drive aneuploidy (Hell et al., 2014). Loss of a functioning VHL gene (located on chromosome arm 3p) has previously been associated with spindle misorientation, chromosome instability and aneuploidy (Thoma et al., 2009). Similarly PBRM1, a gene that encodes BAF180 and also located on chromosome 3p, has been shown to be important for the establishment or maintenance of cohesin on chromatin at centromeres. Loss of functioning PBRM1 has recently been reported as a driver of chromosomal instability and aneuploidy (Brownlee et al., 2014).

The observations in this study suggest that our analysis has been sensitive to known karyotypic patterns and may be useful for the detection of additional patterns.

As well as these commonly cited kidney tissue gene mutations we identified a number of other mutations significantly associated to changes in kideny tissue karyotypes. We identified a number of genes associated with increased overall ploidy change, such as POLE in KIRC tissue and KRAS in KIRP. We found a number of gene mutations associated with the innate immune system and chromosome organisation such as KDM6B, INCENP and H2AFJ which appear to contribute to broad changes in ploidy patterns and we found genes significantly associated with the ploidy pattern signatures we generated in this study which in some cases describe karotypes shared between tissue types.

7 - Discussion

In this thesis I have developed a range of methods to classify and predict potential drug targets that exploit negative genetic interactions such as synthetic lethality or acquired lethality and those associated with genetic instability.

In Chapter 2 I present a review of genetic interactions and SSL, covered the shortcomings of traditional cancer therapies, how therapies that exploit genetic interactions such as SSL might mitigate some of these shortcomings in the next generation of tailored therapies and how drug discovery groups have approached the identification and validation of these interactions. I discussed how one of the primary obstacles to systematic experimental validation of synthetic lethal pairs is currently that of insurmountable experimental burden and how the prediction of SSL pairs through the use of computational methods may help better guide future screens for these genetic interactions. This chapter provided an in-depth literature review and background for the following chapters.

In Chapter 3 I present the Slant algorithm, Slant extends previous attempts to predict human SSL pairs using topological and social features extracted from PPI networks. These models were shown to classify SSL gene pairs with accuracy even when one species SSL pairs were predicted using the training data from another species. These results suggest that many topology patterns associated with SSL gene pairs are conserved between species even if the SSL pairs themselves are not. I demonstrated that network models that focus on pair-wise node features report significantly improved predictive power compared to previous studies that did not utilise pair wise features.

Finally we demonstrated that our models were also more robust than previous studies to pair input bias, a bias introduced when the same genes are present in training and test datasets that feature gene pairs.

Though the classifiers used in Slant demonstrate high predictive power there are some limitations associated with the available source PPI data. These limitations include incompleteness, a lack of directionality and a lack of functional information for each interaction, for example does an interaction result in an increase or decrease in the target protein's activity. Some of the predictive power lost due to to incompleteness was quantified in the study and our model was fairly robust to this incompleteness. I note that similar models should report increased accuracy as the known PPI network is improved through more sophisticated screening. Related to these limitations the pairs predicted in the Slant study are based on a generic, consensus version of the PPI network which may not be fully representative of the PPI networks of all cancer cell-lines. As such the potential for personalised therapies is limited.

In Chapter 4 I introduced Slorth, a publicly available online database developed to allow researchers and clinicians to easily browse and search for clinically relevant SSL pairs. Slorth features SSL predictions produced via the classifiers introduced in Chapter 3 as well as experimentally validated SSL pairs from BioGRID. The overriding motivation for the Slorth database is to provide public and easy access to high quality SSL predictions to guide future screening.

In Chapter 5 I further developed the idea of using network topology features to classify potential therapy targets, this time extending the idea into genuinely personalised medicine. By modifying PPI networks based on the respective patient's unique genetic alteration profile I was able to model individual cell lines which were in turn used to predict genes that had acquired essentiality in the given cancer patient. Individual PPI networks were modified by removing nodes associated with loss of function mutations in the patient and modulating edge weight based on gain of function mutations and gene expression changes. Again these classifiers reported predictive power both when classifying dependency genes within a cell-line and across cell-lines. This suggests that patterns of PPI associated with acquired essential genes is conserved in a similar way to that observed in the cross-species classification of SSL.

The incompleteness of the known PPI as well as lack of directionality and functional outcome of interactions was again a limiting factor in this study. To fairly model edge weights without directional information all edges were made bi-directional with individual weights based on the expression level and mutational status of the associated source node. This model is unlikely to accurately represent the full complexity of the real interactome and again improvements in the base data will offer further opportunities to improve this type of study.

Despite these limitations this study presents a novel way of personalising an otherwise generic biological network and furthermore reports success when using these models to predict potential drug targets.

Finally in Chapter 6 I investigated how patterns of ploidy could be used as an alternative approach to identifying genes that were associated with genetic instability and cancer progression. I describe patterns of ploidy across chromosome arms including correlation of gains and losses and ploidy signatures using NMF. These patterns matched with previously described changes in kidney cancer issues and highlighted a number of previously undescribed patterns. Using these patterns we identified a range of genes that may be associated with genetic stability and as such could lead to novel drug targets.

An overarching theme of this work has been to add additional value to existing publicly available biological data through the development of novel models that utilise data from a range of sources. These models have led to a SSL classifier that is robust to pair input bias, the public availability of nearly one million high quality SSL and SDL predictions, each a potential therapy target, and a novel approach to predicting personalised cancer drug targets.

Despite the scope for improvement the known PPI network has proved a valuable and flexible resource for building models with the purpose of drug target classification. The modification of this resource to better model individuals or groups is especially promising and may present many opportunities for personalised medicine.

8 - References

A Ari Hakimi, Irina Ostrovnaya, Boris Reva, Nikolaus Schultz, Ying-Bei Chen, Mithat Gonen, Han Liu, Shugaku Takeda, Martin H Voss, Satish K Tickoo, Victor E Reuter, Paul Russo, Emily H. Cheng, Chris Sander, R., and rt J. Motzer, and J.J.H. (2013). Adverse Outcomes in Clear Cell Renal Cell Carcinoma with Mutations of 3p21 Epigenetic Regulators BAP1 and SETD2: a Report by MSKCC and the KIRC TCGA Research NetworkDegennaro, Matthew Hurd, Thomas Ryan Siekhaus, Daria Elisabeth Biteau, Benoit Jasper, Hein. Clin. Cancer Res. *20*, 233–243.

Abbotts, R., Jewell, R., Nsengimana, J., Maloney, D.J., Simeonov, A., Seedhouse, C., Elliott, F., Laye, J., Walker, C., Jadhav, A., et al. (2014). Targeting human apurinic/apyrimidinic endonuclease 1 (APE1) in phosphatase and tensin homolog (PTEN) deficient melanoma cells for personalized therapy. Oncotarget *5*, 3273–3286.

Abdollahpouri, H., Burke, R., and Mobasher, B. (2017). Controlling Popularity Bias in Learning-to-Rank Recommendation. In Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys ’17, pp. 42–46.

Acencio, M.L.M.M.L., Lemke, N., Kobayashi, K., Ehrlich, S., Albertini, A., Amati, G., Andersen, K., Arnaud, M., Asai, K., Ashikaga, S., et al. (2009). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics *10*, 290.

Advani, A.S., and Pendergast, A.M. (2002). Bcr-Abl variants: Biological and clinical aspects. Leuk. Res.

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods.

Aguilar-Quesada, R., Muñoz-Gámez, J.A., Martín-Oliva, D., Peralta, A., Valenzuela, M.T., Matínez-Romero, R., Quiles-Pérez, R., Menissier-de Murcia, J., de Murcia, G., Ruiz de Almodóvar, M., et al. (2007). Interaction between ATM and PARP-1 in response to DNA damage and sensitization of ATM deficient cells through PARP inhibition. BMC Mol. Biol. *8*.

Aguirre, A.J., Meyers, R.M., Weir, B.A., Vazquez, F., Zhang, C.Z., Ben-David, U., Cook, A., Ha, G., Harrington, W.F., Doshi, M.B., et al. (2016). Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. Cancer Discov. *6*, 914–929.

Aksoy, B.A., Demir, E., Babur, Ö., Wang, W., Jing, X., Schultz, N., and Sander, C. (2014). Prediction of individualized therapeutic vulnerabilities in cancer from genomic profiles. Bioinformatics *30*, 2051–2059.

Almaas, E. (2007). Biological impacts and context of network theory. J. Exp. Biol.

Almal, S.H., and Padh, H. (2012). Implications of gene copy-number variation in health and diseases. J. Hum. Genet.

Amaral, L.A.N. (2008). A truer measure of our ignorance. Proc. Natl. Acad. Sci. U. S. A. *105*, 6795–6796.

Ames, B.N., and Gold, L.S. (1991). Endogenous mutagens and the causes of aging and cancer. Mutat. Res. - Fundam. Mol. Mech. Mutagen.

Ames, B.N., Durston, W.E., Yamasaki, E., and Lee, F.D. (1973). Carcinogens are mutagens: a simple test combining liver homogenates for activation and bacteria for detection. Proc. Natl. Acad. Sci. U. S. A.

Anderson, M.W., Reynolds, S.H., You, M., and Maronpot, R.M. (1992). Role of Proto-oncogene activation in carcinogenesis. In Environmental Health Perspectives, p.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol.

Aparicio, T., Baer, R., and Gautier, J. (2014). DNA double-strand break repair pathway choice and cancer. DNA Repair (Amst).

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: Tool for the unification of biology. Nat. Genet. *25*, 25–29.

Asiful Islam, M., Alam, F., Kamal, M.A., Gan, S.H., Wong, K.K., and Sasongko, T.H. (2017). Therapeutic Suppression of Nonsense Mutation: An Emerging Target in Multiple Diseases and Thrombotic Disorders. Curr. Pharm. Des.

Bächle, M., and Kirchberg, P. (2007). Ruby on rails. IEEE Softw.

Bader, G.D., and Hogue, C.W. V (2003). BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res.

Baeissa, H. (2019). Identifying driver mutations in cancers.

Baeissa, H., Benstead-Hume, G., Richardson, C.J.C.J., and Pearl, F.M.G.F.M.G. (2017a). Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. Oncotarget *8*, 21290–21304.

Baeissa, H., Benstead-Hume, G., Richardson, C.J., and Pearl, F.M.G. (2017b). Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. Oncotarget *8*, 21290–21304.

Bagci, O., and Kurtgöz, S. (2015). Amplification of cellular oncogenes in solid tumors. N. Am. J. Med. Sci.

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.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R., et al. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br. J. Cancer *2*, 355–358.

Bandyopadhyay, N., Ranka, S., and Kahveci, T. (2011). SSLpred: Predicting synthetic sickness lethality. p.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112.

Bartha, I., di Iulio, J., Venter, J.C., and Telenti, A. (2018). Human gene essentiality. Nat. Rev. Genet. *19*, 51–62.

Batra, J., Srinivasan, S., and Clements, J. (2014). Single nucleotide polymorphisms (SNPs). In Molecular Testing in Cancer, p.

Behjati, S., Gundem, G., Wedge, D.C., Roberts, N.D., Tarpey, P.S., Cooke, S.L., Van Loo, P., Alexandrov, L.B., Ramakrishna, M., Davies, H., et al. (2016). Mutational signatures of ionizing radiation in second malignancies. Nat. Commun.

Bengio, Y., and Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. J. Mach. Learn. Res.

Benstead-Hume, G., Wooller, S.K.S.K., and Pearl, F.M.G.F.M.G. (2017a). Computational Approaches to Identify Genetic Interactions for Cancer Therapeutics. J. Integr. Bioinform. *14*, 1–12.

Benstead-Hume, G., Wooller, S.K., and Pearl, F.M.G. (2017b). ‘Big data’ approaches for novel anti-cancer drug discovery. Expert Opin. Drug Discov. *12*.

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.

Berns, K., Hijmans, E.M., Mullenders, J., Brummelkamp, T.R., Velds, A., Heimerikx, M., Kerkhoven, R.M., Madlredjo, M., Nijkamp, W., Weigelt, B., et al. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature.

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.

Bitler, B.G., Aird, K.M., Garipov, A., Li, H., Amatangelo, M., Kossenkov, A. V., Schultz, D.C., Liu, Q., Shih, I.-M.M., Conejo-Garcia, J.R., et al. (2015). Synthetic lethality by targeting EZH2 methyltransferase activity in ARID1A-mutated cancers. Nat. Med.

Blot, W.J., McLaughlin, J.K., Winn, D.M., Austin, D.F., Greenberg, R.S., Susan, S., Preston, M., Bernstein, L., Schoenberg, J.B., Stemhagen, A., et al. (1988). Smoking and Drinking in Relation to Oral and Pharyngeal Cancer. Cancer Res.

Boucher, B., and Jenna, S. (2013). Genetic interaction networks: Better understand to better predict. Front. Genet. *4*, 1–16.

Bouck, N., Stellmach, V., and Hsu, S.C. (1996). How Tumors Become Angiogenic. p.

Breitkreutz, B.J., Stark, C., and Tyers, M. (2002). The GRID: The General Repository for Interaction Datasets. Genome Biol.

Brownlee, J. (2016). Overfitting and Underfitting With Machine Learning Algorithms.

Brownlee, P.M., Chambers, A.L., Oliver, A.W., and Downs, J.A. (2012). Cancer and the bromodomains of BAF180. Biochem. Soc. Trans. *40*, 364–369.

Brownlee, P.M., Chambers, A.L., Cloney, R., Bianchi, A., and Downs, J. a. (2014). BAF180 Promotes Cohesion and Prevents Genome Instability and Aneuploidy. Cell Rep. *6*, 973–981.

Bryan, T.M., and Cech, T.R. (1999). Telomerase and the maintenance of chromosome ends. Curr. Opin. Cell Biol.

Bryant, H.E., Schultz, N., Thomas, H.D., Parker, K.M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N.J., and Helleday, T. (2005). Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature *434*, 913–917.

Buehler, E., Chen, Y.C., and Martin, S. (2012). C911: A Bench-Level Control for Sequence Specific siRNA Off-Target Effects. PLoS One *7*.

Bulusu, K.C., Tym, J.E., Coker, E.A., Schierz, A.C., and Al-Lazikani, B. (2014). CanSAR: Updated cancer research and drug discovery knowledgebase. Nucleic Acids Res. *42*, 61–63.

Bunaciu, A.A., Udriştioiu, E. gabriela, and Aboul-Enein, H.Y. (2015). X-Ray Diffraction: Instrumentation and Applications. Crit. Rev. Anal. Chem.

Bunting, S.F., and Nussenzweig, A. (2013). End-joining, translocations and cancer. Nat. Rev. Cancer.

Burgess, A., Chia, K.M., Haupt, S., Thomas, D., Haupt, Y., and Lim, E. (2016). Clinical Overview of MDM2/X-Targeted Therapies. Front. Oncol. *6*, 7.

Burrell, R.A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. Nature.

Byrne, A.B., Weirauch, M.T., Wong, V., Koeva, M., Dixon, S.J., Stuart, J.M., and Roy, P.J. (2007). A global analysis of genetic interactions in Caenorhabditis elegans. J. Biol. *6*, 8.

Cancer, T., and Line, C. (2015). Pharmacogenomic agreement between two cancer cell line data sets. Nature *528*, 84–87.

Canese, K., and Weis, S. (2013). PubMed: The bibliographic database. NCBI Handb.

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.

Chan, D.A., and Giaccia, A.J. (2011). Harnessing synthetic lethal interactions in anticancer drug discovery. Nat. Rev. Drug Discov. *10*, 351–364.

Charlton, P., and Spicer, J. (2016). Targeted therapy in cancer. Med. (United Kingdom).

Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al. (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res. *43*, D470-8.

Chen, Y., and Xu, D. (2005). Understanding protein dispensability through machine-learning analysis of high-throughput data. Bioinformatics *21*, 575–581.

Chen, F., Zhang, Y., Şenbabaoğlu, Y., Ciriello, G., Yang, L., Reznik, E., Shuch, B., Micevic, G., De Velasco, G., Shinbrot, E., et al. (2016). Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. Cell Rep. 2476–2489.

Cheng, J., Xu, Z., Wu, W., Zhao, L., Li, X., Liu, Y., and Tao, S. (2014). Training set selection for the prediction of essential genes. PLoS One *9*.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., et al. (1998). SGD: Saccharomyces genome database. Nucleic Acids Res.

Cheung, H.W., Cowley, G.S., Weir, B.A., Boehm, J.S., Rusin, S., Scott, J.A., East, A., Ali, L.D., Lizotte, P.H., Wong, T.C., et al. (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc. Natl. Acad. Sci. *108*, 12372–12377.

Chipman, K.C., and Singh, A.K. (2009). Predicting genetic interactions with random walks on biological networks. BMC Bioinformatics *10*, 17.

Cho, H., Berger, B., and Peng, J. (2016). Compact Integration of Multi-Network Topology for Functional Analysis of Genes. Cell Syst. *3*, 540-548.e5.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. *22*, 398–406.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. *45*, 1127–1133.

Coates, A., Abraham, S., Kaye, S.B., Sowerbutts, T., Frewin, C., Fox, R.M., and Tattersall, M.H.N. (1983). On the receiving end-patient perception of the side-effects of cancer chemotherapy. Eur. J. Cancer Clin. Oncol. *19*, 203–208.

Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M., et al. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature *446*, 806–810.

Collins, S.R., Roguev, A., and Krogan, N.J. (2010). Quantitative genetic interaction mapping using the E-Map approach. Methods Enzymol.

Conde-Pueyo, N., Munteanu, A., Solé, R. V., and Rodríguez-Caso, C. (2009). Human synthetic lethal inference as potential anti-cancer target gene detection. BMC Syst. Biol.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature.

Costa-Cabral, S., Brough, R., Konde, A., Aarts, M., Campbell, J., Marinari, E., Riffell, J., Bardelli, A., Torrance, C., Lord, C.J., et al. (2016). CDK1 Is a Synthetic Lethal Target for KRAS Mutant Tumours. PLoS One *11*, e0149099.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science (80-.). *327*, 425–431.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M.M.M.M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science (80-.). *353*, aaf1420–aaf1420.

Cowley, G.S., Weir, B.A., Vazquez, F., Tamayo, P., Scott, J.A., Rusin, S., East-Seletsky, A., Ali, L.D., Gerath, W.F.J., Pantel, S.E., et al. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci. Data *1*.

Croce, C.M. (2008). Oncogenes and cancer. N. Engl. J. Med.

Csárdi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Syst. *1695*, 1–9.

D'Antonio, M., Guerra, R.F., Cereda, M., Marchesi, S., Montani, F., Nicassio, F., DiFiore, P.P., and Ciccarelli, F.D. (2013). Recessive cancer genes engage in negative genetic interactions with their functional paralogs. Cell Rep. *5*, 1519–1526.

D1Etterich, T. (1995). Overfitting and Undercomputing in Machine Learning. ACM Comput. Surv.

Dallago, C., Goldberg, T., Andrade-Navarro, M.A., Alanis-Lobato, G., and Rost, B. (2018). CellMap visualizes protein-protein interactions and subcellular localization. F1000Research.

Dang, C. V. (2012). MYC on the path to cancer. Cell.

Datto, M.B., Hu, P.P., Kowalik, T.F., Yingling, J., and Wang, X.F. (1997). The viral oncoprotein E1A blocks transforming growth factor beta-mediated induction of p21/WAF1/Cip1 and p15/INK4B. Mol. Cell. Biol.

Degtyarenko, K., De matos, P., Ennis, M., Hastings, J., Zbinden, M., Mcnaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: A database and ontology for chemical entities of biological interest. Nucleic Acids Res. *36*.

DeMarini, D.M. (2004). Genotoxicity of tobacco smoke and tobacco smoke condensate: A review. Mutat. Res. - Rev. Mutat. Res.

DeNardo, D.G., Andreu, P., and Coussens, L.M. (2010). Interactions between lymphocytes and myeloid cells regulate pro-versus anti-tumor immunity. Cancer Metastasis Rev.

Donehower, L.A., Soussi, T., Korkut, A., Liu, Y., Schultz, A., Cardenas, M., Li, X., Babur, O., Hsu, T.-K., Lichtarge, O., et al. (2019). Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas. Cell Rep.

Douglas Hanahan, R.A.W. (2000). The Hallmarks of Cancer. Cell *100*, 57–70.

Downward, J. (2015). RAS synthetic lethal screens revisited: Still seeking the elusive prize? Clin. Cancer Res.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*.

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res.

Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends Genet.

Elmore, S. (2007). Apoptosis: A Review of Programmed Cell Death. Toxicol. Pathol.

Emerling, B.M., Hurov, J.B., Poulogiannis, G., Choo-Wing, R., Wulf, G.M., Shim, H.-S., Lamia, K.A., Rameh, L.E., Yuan, X., Bullock, A., et al. (2014). Depletion of a putatively druggable class of phosphatidylinositol kinases inhibits growth of p53 null tumors. Cancer Res. *74*, no pagination.

Esplin, E.D., Oei, L., and Snyder, M.P. (2014). Personalized sequencing and the future of medicine: Discovery, diagnosis and defeat of disease. Pharmacogenomics.

Eyfjörd, J.E., Thorlacius, S., Valgardsdottir, R., Gretarsdottir, S., Steinarsdottir, M., and Anamthawat-Jonsson, K. (1995). TP53 abnormalities and genetic instability in breast cancer. Acta Oncol. *34*, 663–667.

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N.J., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature *434*, 917–921.

Fields, S., and Sternglanz, R. (1994). The two-hybrid system: an assay for protein-protein interactions. Trends Genet.

Firth, H. V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Vooren, S. Van, Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet.

Fitzpatrick, T.B., and Sober, A.J. (1985). Sunlight and Skin Cancer. N. Engl. J. Med.

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.

Forbes, S.A., Beare, D., Bindal, N., Bamford, S., Ward, S., Cole, C.G., Jia, M., Kok, C., Boutselakis, H., De, T., et al. (2016). COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. Curr. Protoc. Hum. Genet. *91*, 10.11.1-10.11.37.

Francis, L. (2014). Unsupervised learning. In Predictive Modeling Applications in Actuarial Science: Volume I: Predictive Modeling Techniques, p.

Franken, N.A.P., Rodermond, H.M., Stap, J., Haveman, J., and van Bree, C. (2006). Clonogenic assay of cells in vitro. Nat. Protoc. *1*, 2315–2319.

Fridman, J.S., and Lowe, S.W. (2003). Control of apoptosis by p53. Oncogene.

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.

Gad, H., Koolmeister, T., Jemth, A.-S.S., Eshtad, S., Jacques, S.A., Ström, C.E., Svensson, L.M., Schultz, N., Lundbäck, T., Einarsdottir, B.O., et al. (2014). MTH1 inhibition eradicates cancer by preventing sanitation of the dNTP pool. Nature *508*, 215–221.

Gamudi, D., and Blundell, R. (2010). Tumor suppressor genes. Res. J. Med. Sci.

Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics.

Gelbart, W.M., Rindone, W.P., Chillemi, J., Russo, S., Crosby, M., Mathews, B., Ashburner, M., Drysdale, R.A., De Grey, A., Whitfield, E.J., et al. (1996). FlyBase: The Drosophila database. Nucleic Acids Res. *24*, 53–56.

Geng, L., Zhu, M., Wang, Y., Cheng, Y., Liu, J., Shen, W., Li, Z., Zhang, J., Wang, C., Jin, G., et al. (2016). Genetic variants in chromatin-remodeling pathway associated with lung cancer risk in a Chinese population. Gene *587*, 178–182.

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M.G. Della, Jädersten, M., Dolatshad, H., Verma, A., Cross, N.C.P., Vyas, P., et al. (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nat. Commun. *6*, 5901.

Gilsdorf, M., Horn, T., Arziman, Z., Pelz, O., Kiner, E., and Boutros, M. (2009). GenomeRNAi: A database for cell-based RNAi phenotypes. 2009 update. Nucleic Acids Res. *38*.

Giroux, V., and Rustgi, A.K. (2017). Metaplasia: Tissue injury adaptation and a precursor to the dysplasia-cancer sequence. Nat. Rev. Cancer.

Gnarra, J.R., Tory, K., Weng, Y., Schmidt, L., Wei, M.H., Li, H., Latif, F., Liu, S., Chen, F., Duh, F.M., et al. (1994). Mutations of the VHL tumour suppressor gene in renal carcinoma. Nat. Genet.

Gokul, G., and Khosla, S. (2013). DNA methylation and cancer. Subcell. Biochem.

Goldman, R., and Shields, P.G. (2003). Food Mutagens. J. Nutr.

Gordon, D.J., Resio, B., and Pellman, D. (2012). Causes and consequences of aneuploidy in cancer. Nat. Rev. Genet.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature.

Grivennikov, S.I., Greten, F.R., and Karin, M. (2010). Immunity, Inflammation, and Cancer. Cell.

Guo, J., Liu, H., and Zheng, J. (2015). SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. Nucleic Acids Res. *44*, D1011-1017.

Hanahan, D., and Folkman, J. (1996). Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. Cell.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hannon, G.J., and Beach, D. (1994). PI5INK4B is a potentia| effector of TGF-β-induced cell cycle arrest. Nature.

Hansen, L.L. (2006). Molecular diagnosis of breast cancer. In Prevention and Treatment of Age-Related Diseases, p.

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 (80-.). *278*, 1064–1068.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat. Rev. Genet.

Hell, M.P., Duda, M., Weber, T.C., Moch, H., and Krek, W. (2014). Tumor suppressor VHL functions in the control of mitotic fidelity. Cancer Res. *74*, 2422–2431.

Hengartner, M.O. (2000). The biochemistry of apoptosis. Nature.

Hennequart, M., Pilotte, L., Cane, S., Hoffmann, D., Stroobant, V., Plaen, E. De, and Eynde, B.J. Van den (2017). Constitutive IDO1 Expression in Human Tumors Is Driven by Cyclooxygenase-2 and Mediates Intrinsic Immune Resistance. Cancer Immunol. Res. *5*, 695–709.

Hermjakob, H. (2004). IntAct: an open source molecular interaction database. Nucleic Acids Res.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database.

Hin, A., Tong, Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., et al. (2004). Global Mapping of the Yeast Genetic Interaction Network. Science (80-.). *303*, 808–814.

Hoe, K.K., Verma, C.S., Lane, D.P., Khoo, K.H., Verma, C.S., and Lane, D.P. (2014). Drugging the p53 pathway: understanding the route to clinical efficacy. Nat. Rev. Drug Discov. *13*, 217–236.

Holland, A.J., and Cleveland, D.W. (2012). Losing balance: The origin and impact of aneuploidy in cancer. EMBO Rep.

Holohan, C., Van Schaeybroeck, S., Longley, D.B., and Johnston, P.G. (2013). Cancer drug resistance: An evolving paradigm. Nat. Rev. Cancer *13*, 714–726.

Hopkins, A.L., and Groom, C.R. (2002). The druggable genome. Nat. Rev. Drug Discov. *1*, 727–730.

Hopkins, S.R., McGregor, G.A., Murray, J.M., Downs, J.A., and Savic, V. (2016). Novel synthetic lethality screening method identifies TIP60-dependent radiation sensitivity in the absence of BAF180. DNA Repair (Amst). *46*, 47–54.

Huang, S.X.L., Jaurand, M.C., Kamp, D.W., Whysner, J., and Hei, T.K. (2011). Role of mutagenicity in asbestos fiber-induced carcinogenicity and other diseases. J. Toxicol. Environ. Heal. - Part B Crit. Rev.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.

Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. Nature *545*, 505–509.

Huxford, T. (2013). X-Ray Crystallography. In Brenner’s Encyclopedia of Genetics: Second Edition, p.

Hwang, Y.C., Lin, C.C., Chang, J.Y., Mori, H., Juan, H.F., and Huang, H.C. (2009). Predicting essential genes based on network and sequence analysis. Mol. Biosyst. *5*, 1672–1678.

Iguyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. J. Mach. Learn. Res.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell *166*, 740–754.

Iorns, E., Lord, C.J., Turner, N., and Ashworth, A. (2007). Utilizing RNA interference to enhance cancer drug discovery. Nat. Rev. Drug Discov. *6*, 556–568.

Jackson, A.L., and Linsley, P.S. (2004). Noise amidst the silence: Off-target effects of siRNAs? Trends Genet. *20*, 521–524.

Jackson, S.P., and Bartek, J. (2009). The DNA-damage response in human biology and disease. Nature.

Jacunski, A., Dixon, S.J., and Tatonetti, N.P. (2015). Connectivity Homology Enables Inter-Species Network Models of Synthetic Lethality. PLoS Comput. Biol. *11*, 1–29.

James Kent, W., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res.

Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.

Jerby-Arnon, L., Pfetzer, N., Waldman, Y.Y.Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P.A.A., et al. (2014). Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality. Cell *158*, 1199–1209.

Jiang, W.G., Sanders, A.J., Katoh, M., Ungefroren, H., Gieseler, F., Prince, M., Thompson, S.K., Zollo, M., Spano, D., Dhawan, P., et al. (2015). Tissue invasion and metastasis: Molecular, biological and clinical perspectives. Semin. Cancer Biol.

Jones, P.A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. Nat. Rev. Genet.

Jonsson, P.F., and Bates, P.A. (2006). Global topological features of cancer proteins in the human interactome. Bioinformatics.

Kaelin Jr, W.G., and Kaelin, W.G. (2005). The concept of synthetic lethality in the context of anticancer therapy. Nat. Rev. Cancer *5*, 689–698.

Kaiser, M.F., Johnson, D.C., Wu, P., Walker, B.A., Brioli, A., Mirabella, F., Wardell, C.P., Melchor, L., Davies, F.E., and Morgan, G.J. (2013). Global methylation analysis identifies prognostically important epigenetically inactivated tumor suppressor genes in multiple myeloma. Blood.

Kalbitzer, H.R. (1999). Protein NMR Techniques. Zeitschrift Für Phys. Chemie.

Kamata, T., Hussain, J., Giblett, S., Hayward, R., Marais, R., and Pritchard, C. (2010). BRAf inactivation drives aneuploidy by deregulating CRAF. Cancer Res. *70*, 8475–8486.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature.

Karnitz, L.M., and Zou, L. (2015). Molecular pathways: Targeting ATR in cancer therapy. Clin. Cancer Res. *21*, 4780–4785.

Karnoub, A.E., and Weinberg, R.A. (2016). Chemokine Networks and Breast Cancer Metastasis. Breast Dis.

Karp, P.D. (2002). The MetaCyc Database. Nucleic Acids Res. *30*, 59–61.

Kelley, R., and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. Nat. Biotechnol. *23*, 561–566.

Kim, T., Xi, R., Luquette, L.J., Park, R.W., Johnson, M.D., and Park, P.J. (2013). Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. 217–227.

Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). *2011*, bar030.

Klein, C.A. (2008). Cancer: The metastasis cascade. Science (80-.).

Klikauer, T. (2016). Scikit-learn: Machine Learning in Python.

Kotsiantis, S.B. (2007). Supervised machine learning: A review of classification techniques. Inform.

Kranthi, T., Rao, S.B., and Manimaran, P. (2013). Identification of synthetic lethal pairs in biological systems through network information centrality. Mol. Biosyst. *9*, 2163–2167.

Krause, S.A., and Gray, J. V (2009). The functional relationships underlying a synthetic genetic network. Commun Integr Biol *2*, 4–6.

Kumar, M.S., Hancock, D.C., Molina-Arcas, M., Steckel, M., East, P., Diefenbacher, M., Armenteros-Monterroso, E., Lassailly, F., Matthews, N., Nye, E., et al. (2012). The

GATA2 transcriptional network is requisite for RAS oncogene-driven non-small cell lung cancer. Cell.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). SIFT. Nat. Protoc. *4*, 1073–1081.

Kushi, L., and Giovannucci, E. (2002). Dietary fat and cancer. In American Journal of Medicine, p.

Laddha, S. V, Ganesan, S., Chan, C.S., and White, E. (2014). Mutational landscape of the essential autophagy gene BECN1 in human cancers. Mol. Cancer Res. *12*, 485–490.

De Las Rivas, J., and Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput. Biol. *6*, e1000807.

Lazar, C., Nuzillard, D., Billaudel, P., and Curila, S. (2009). Non negative matrix factorization clustering capabilities; application on multivariate image segmentation. J. Electr. Electron. Eng.

Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems, p.

Lehman, T.A., Reddel, R., Pfeifer, A.M.A., Spillare, E., Kaighn, M.E., Weston, A., Gerwin, B.I., and Harris, C.C. (1991). Oncogenes and tumor-suppressor genes. In Environmental Health Perspectives, p.

Lehne, B., and Schlitt, T. (2009). Protein-protein interaction databases: keeping up with growing interactomes. Hum. Genomics.

Leiserson, M.D.M., Wu, H.T., Vandin, F., and Raphael, B.J. (2015). CoMEt: A statistical approach to identify combinations of mutually exclusive alterations in cancer. Genome Biol.

Li, B., Cao, W., Zhou, J., and Luo, F. (2011). Understanding and predicting synthetic lethal genetic interactions in Saccharomyces cerevisiae using domain genetic interactions. BMC Syst. Biol. *5*, 73.

Li, L., Zhang, K., Lee, J., Cordes, S., Davis, D.P., and Tang, Z. (2009). Discovering cancer genes by integrating network and functional properties. BMC Med. Genomics.

Li, L., Shen, C., Nakamura, E., Ando, K., Signoretti, S., Beroukhim, R., Cowley, G.S., Lizotte, P., Liberzon, E., Bair, S., et al. (2013). SQSTM1 Is a Pathogenic Target of 5q Copy Number Gains in Kidney Cancer. Cancer Cell.

Li, X., Mishra, S.K., Wu, M., Zhang, F., and Zheng, J. (2014). Syn-lethality: an integrative knowledge base of synthetic lethality towards discovery of selective anticancer therapies. Biomed Res. Int. *2014*, 196034.

Lindeman, N.I., Cagle, P.T., Beasley, M.B., Chitale, D.A., Dacic, S., Giaccone, G., Jenkins, R.B., Kwiatkowski, D.J., Saldivar, J.-S., Squire, J., et al. (2013). Molecular Testing Guideline for Selection of Lung Cancer Patients for EGFR and ALK Tyrosine Kinase Inhibitors. J. Mol. Diagnostics *15*, 415–453.

Linding, R., Jensen, L.J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M.B., and Pawson, T. (2008). NetworKIN: A resource for exploring cellular phosphorylation networks. Nucleic Acids Res.

Litton, J.K.K., Scoggins, M., Ramirez, D.L.L., Murthy, R.K.K., Whitman, G.J.J., Hess, K.R.R., Adrada, B.E.E., Moulder, S.L.L., Barcenas, C.H.H., Valero, V., et al. (2017). A pilot study of neoadjuvant talazoparib for early-stage breast cancer patients with a BRCA mutation. Ann. Oncol. *27*, 153PD.

Liu, J.F., Konstantinopoulos, P.A., and Matulonis, U.A. (2014). PARP inhibitors in ovarian cancer: current status and future promise. Gynecol. Oncol. *133*, 362–369.

Lord, C.J., and Ashworth, A. (2012). The DNA damage response and cancer therapy. Nature *481*, 287–294.

Lord, C.J., and Ashworth, A. (2017). PARP inhibitors: Synthetic lethality in the clinic. Science (80-.).

Lu, X., Kensche, P.R., Huynen, M.A., and Notebaart, R.A. (2013). Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. Nat. Commun.

Lu, Y., Deng, J., Rhodes, J.C., Lu, H., and Lu, L.J. (2014). Predicting essential genes for identifying potential drug targets in Aspergillus fumigatus. Comput. Biol. Chem. *50*, 29–40.

Lunt, S.Y., and Vander Heiden, M.G. (2011). Aerobic Glycolysis: Meeting the Metabolic Requirements of Cell Proliferation. Annu. Rev. Cell Dev. Biol.

Luo, B., Cheung, H.W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J.S., Beroukhim, R., Weir, B.A., et al. (2008). Highly parallel identification of essential genes in cancer cells. Proc. Natl. Acad. Sci. U. S. A. *105*, 20380–20385.

Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.-K.K., and Elledge, S.J. (2009a). A Genome-wide RNAi Screen Identifies Multiple Synthetic Lethal Interactions with the Ras Oncogene. Cell *137*, 835–848.

Luo, J., Solimini, N.L., and Elledge, S.J. (2009b). Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction. Cell *136*, 823–837.

Lynch, D.H. (1987). Oncogenes and cancer. Am. J. Reprod. Immunol. Microbiol.

Ma, J., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods *15*, 290–298.

Madhukar, N.S., Elemento, O., and Pandey, G. (2015a). Prediction of genetic interactions using machine learning and network properties. Front. Bioeng. Biotechnol.

Madhukar, N.S., Elemento, O., and Pandey, G. (2015b). Prediction of Genetic Interactions Using Machine Learning and Network Properties. Front. Bioeng. Biotechnol. *3*, 172.

Manco, G., Ritacco, E., and Guarascio, M. (2019). Network Topology. In Encyclopedia of Bioinformatics and Computational Biology, p.

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.

Marcotte, R., Sayad, A., Brown, K.R., Pe, D., Moffat, J., Neel, B.G., Drivers, C., Marcotte, R., Sayad, A., Brown, K.R., et al. (2016). Functional Genomic Landscape of Human Breast Resource Functional Genomic Landscape of Human Breast Cancer Drivers , Vulnerabilities , and Resistance. Cell 293–309.

Martin, Z. (2016). Rules of Machine Learning : Best Practices for ML Engineering. Google.

Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science (80-.).

Mateo, J., Carreira, S., Sandhu, S., Miranda, S., Mossop, H., Perez-Lopez, R., Rodrigues, D.N., Robinson, D., Omlin, A., Tunariu, N., et al. (2015). DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. N. Engl. J. Med. *368*, 2255–2265.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, p.

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.

Megchelenbrink, W., Katzir, R., Lu, X., Ruppin, E., and Notebaart, R.A. (2015). Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient survival. Proc. Natl. Acad. Sci. U. S. A. *112*, 12217–12222.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F.A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. *33*.

Mesri, E.A., Feitelson, M.A., and Munger, K. (2014). Human viral oncogenesis: A cancer hallmarks analysis. Cell Host Microbe.

Michaut, M., and Bader, G.D. (2012). Multiple genetic interaction experiments provide complementary information useful for gene function prediction. PLoS Comput. Biol. *8*.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science (80-.).

Miller, C.A., Settle, S.H., Sulman, E.P., Aldape, K.D., and Milosavljevic, A. (2011). Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Med. Genomics.

Millet, C., and Makovets, S. (2016). Aneuploidy as a mechanism of adaptation to telomerase insufficiency. Curr. Genet.

Mosca, R., Pons, T., Céol, A., Valencia, A., and Aloy, P. (2013). Towards a detailed atlas of protein-protein interactions. Curr. Opin. Struct. Biol. *23*, 929–940.

Muller, P.Y., and Milton, M.N. (2012). The determination and interpretation of the therapeutic index in drug development. Nat. Rev. Drug Discov. *11*, 751–761.

Muller, F.L., Colla, S., Aquilanti, E., Manzo, V.E., Genovese, G., Lee, J., Eisenson, D., Narurkar, R., Deng, P., Nezi, L., et al. (2012). Passenger deletions generate therapeutic vulnerabilities in cancer. Nature *488*, 337–342.

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.

Nagahashi, M., Sato, S., Yuza, K., Shimada, Y., Ichikawa, H., Watanabe, S., Takada, K., Okamoto, T., Okuda, S., Lyle, S., et al. (2018). Common driver mutations and smoking history affect tumor mutation burden in lung adenocarcinoma. J. Surg. Res.

Nambiar, M., Kari, V., and Raghavan, S.C. (2008). Chromosomal translocations in cancer. Biochim. Biophys. Acta - Rev. Cancer.

National Cancer Institute (2019). Genomic Data Commons Data Portal.

Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol.

Ngo, V.N., Davis, R.E., Lamy, L., Yu, X., Zhao, H., Lenz, G., Lam, L.T., Dave, S., Yang, L., Powell, J., et al. (2006). A loss-of-function RNA interference screen for molecular targets in cancer. Nature *441*, 106–110.

Nguyen, D.T., Mathias, S., Bologa, C., Brunak, S., Fernandez, N., Gaulton, A., Hersey, A., Holmes, J., Jensen, L.J., Karlsson, A., et al. (2017). Collating protein information to shed light on the druggable genome. Genome Biol. Evol. *45*, D995–D1002.

Nijman, S.M.B.B. (2011). Synthetic lethality: General principles, utility and detection using genetic screens in human cells. FEBS Lett. *585*, 1–6.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979–993.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *27*, 29–34.

Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and Kinoshita, K. (2015). COXPRESdb in 2015: Coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. Nucleic Acids Res.

Ooi, S.L., Shoemaker, D.D., and Boeke, J.D. (2003). DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. Nat. Genet. *35*, 277–286.

Orr, B., Godek, K.M., and Compton, D. (2015). Aneuploidy. Curr. Biol. *25*, R538.

Oshiro, T.M., Perez, P.S., and Baranauskas, J.A. (2012). How many trees in a random forest? In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), p.

Paladugu, S.R., Zhao, S., Ray, A., and Raval, A. (2008). Mining protein networks for synthetic genetic interactions. BMC Bioinformatics *9*, 426.

Pan, X., Yuan, D.S., Ooi, S.L., Wang, X., Sookhai-Mahadeo, S., Meluh, P., and Boeke, J.D. (2007). dSLAM analysis of genome-wide genetic interactions in Saccharomyces cerevisiae. Methods.

Papke, B., and Der, C.J. (2017). Drugging RAS: Know the enemy. Science (80-.).

Park, C.-K., and Kim, D.G. (2012). K-FOLD Cross-Validation. Prog. Neurol. Surg.

Park, Y., and Marcotte, E.M. (2012). Flaws in evaluation schemes for pair-input computational predictions. Nat. Methods *9*, 1134–1136.

Pavlopoulos, G.A., Secrier, M., Moschopoulos, C.N., Soldatos, T.G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P.G. (2011). Using graph theory to analyze biological networks. BioData Min.

Pearl, L.H., Schierz, A.C., Ward, S.E., Al-Lazikani, B., and Pearl, F.M.G.G. (2015). Therapeutic opportunities within the DNA damage response. Nat. Rev. Cancer *15*, 166–180.

Peri, S. (2004). Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res.

Pesquita, C., Faria, D., Falcão, A.O., Lord, P., and Couto, F.M. (2009). Semantic similarity in biomedical ontologies. PLoS Comput. Biol. *5*.

Petryszak, R., Keays, M., Tang, Y.A., Fonseca, N.A., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A.M.P., Jupp, S., Koskinen, S., et al. (2016). Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res.

Plaimas, K., Eils, R., and König, R. (2010). Identifying essential genes in bacterial metabolic networks with machine learning methods. BMC Syst. Biol. *4*.

Pochapsky, T.C., and Pochapsky, S.S. (2013). Nuclear magnetic resonance spectroscopy. In Molecular Biophysics for the Life Sciences, p.

Pon, J.R., and Marra, M.A. (2015). Driver and Passenger Mutations in Cancer. Annu. Rev. Pathol. Mech. Dis.

Porcelli, L., E. Quatrale, A., Mantuano, P., Silvestris, N., E. Brunetti, A., Calvert, H., Paradiso, A., and Azzariti, A. (2012). Synthetic Lethality to Overcome Cancer Drug Resistance. Curr. Med. Chem. *19*, 3858–3873.

Puc, J., Keniry, M., Li, H.S., Pandita, T.K., Choudhury, A.D., Memeo, L., Mansukhani, M., Murty, V.V.V.S., Gaciong, Z., Meek, S.E.M., et al. (2005). Lack of PTEN sequesters CHK1 and initiates genetic instability. Cancer Cell *7*, 193–204.

Puyol, M., Martín, A., Dubus, P., Mulero, F., Pizcueta, P., Khan, G., Guerra, C., Santamaría, D., and Barbacid, M. (2010). A Synthetic Lethal Interaction between K-Ras Oncogenes and Cdk4 Unveils a Therapeutic Strategy for Non-small Cell Lung Carcinoma. Cancer Cell *18*, 63–73.

Pylayeva-Gupta, Y., Grabocka, E., and Bar-Sagi, D. (2011). RAS oncogenes: Weaving a tumorigenic web. Nat. Rev. Cancer *11*, 761–774.

Qi, Y. (2012). Random forest for bioinformatics. In Ensemble Machine Learning: Methods and ApplicatiOns, p.

Qian, B.Z., and Pollard, J.W. (2010). Macrophage Diversity Enhances Tumor Progression and Metastasis. Cell.

R Development Core Team, R. (2011). R: A Language and Environment for Statistical Computing.

Rajagopalan, H., and Lengauer, C. (2004). Aneuploidy and cancer. Nature.

Rekhadevi, P. V., Mahboob, M., Rahman, M.F., and Grover, P. (2009). Genetic damage in wood dust-exposed workers. Mutagenesis.

Reva, B.A., Antipin, Y.A. and Sander, C. (2010). Mutation Assessor. Cancer.

Richardson, C.J., Gao, Q., Mitsopoulous, C., Zvelebil, M., Pearl, L.H., and Pearl, F.M.G. (2009). MoKCa database - Mutations of kinases in cancer. Nucleic Acids Res.

Riley, L.B., and Anderson, D.W. (2011). Cancer epigenetics. In Handbook of Epigenetics, p.

Rodríguez-Paredes, M., and Esteller, M. (2011). Cancer epigenetics reaches mainstream oncology. Nat. Med.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

Rouas-Freiss, N., Moreau, P., Menier, C., and Carosella, E.D. (2003). HLA-G in cancer: A way to turn off the immune system. Semin. Cancer Biol.

Roumeliotis, T.I., Williams, S.P., Gonçalves, E., Alsinet, C., Del Castillo Velasco-Herrera, M., Aben, N., Ghavidel, F.Z., Michaut, M., Schubert, M., Price, S., et al. (2017). Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells. Cell Rep. *20*, 2201–2214.

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M., and Sali, A. (2004). A structural perspective on protein-protein interactions. Curr. Opin. Struct. Biol.

Ryan, C.J., Lord, C.J., and Ashworth, A. (2014). DAISY: Picking synthetic lethals from cancer genomes. Cancer Cell *26*, 306–308.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics.

Saha, S., and Heber, S. (2006). In silico prediction of yeast deletion phenotypes. Genet. Mol. Res. [Electronic Resour. GMR. *5*, 224–232.

Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I., et al. (2016). A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. *16*, 19–34.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: The pathway interaction database. Nucleic Acids Res.

Scholl, C., Fröhling, S., Dunn, I.F., Schinzel, A.C., Barbie, D.A., Kim, S.Y., Silver, S.J., Tamayo, P., Wadlow, R.C., Ramaswamy, S., et al. (2009). Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. Cell *137*, 821–834.

Schou, C., and Heegaard, N.H.H. (2006). Recent applications of affinity interactions in capillary electrophoresis. Electrophoresis.

Schrank, Z., Chhabra, G., Lin, L., Iderzorig, T., Osude, C., Khan, N., Kuckovic, A., Singh, S., Miller, R.J., and Puri, N. (2018). Current molecular-targeted therapies in NSCLC and their mechanism of resistance. Cancers (Basel).

Schwab, M. (1999). Oncogene amplification in solid tumors. Semin. Cancer Biol.

Segal, M.R. (2004). Machine Learning Benchmarks and Random Forest Regression. Biostatistics.

Sehn, J.K. (2014). Insertions and Deletions (Indels). In Clinical Genomics, p.

Sharma, S., Kelly, T.K., and Jones, P.A. (2009). Epigenetics in cancer. Carcinogenesis.

Shawver, L.K., Slamon, D., and Ullrich, A. (2002). Smart drugs: Tyrosine kinase inhibitors in cancer therapy. Cancer Cell *1*, 117–123.

Shay, J.W., and Bacchetti, S. (1997). A survey of telomerase activity in human cancer. Eur. J. Cancer Part A.

Shen, J., Peng, Y., Wei, L., Zhang, W., Yang, L., Lan, L., Kapoor, P., Ju, Z., Mo, Q., Shih, I.M., et al. (2015). ARID1A Deficiency Impairs the DNA Damage Checkpoint and Sensitizes Cells to PARP Inhibitors. Cancer Discov. *5*, 752–767.

Shepherd, F.A., Rodrigues Pereira, J., Ciuleanu, T., Tan, E.H., Hirsh, V., Thongprasert, S., Campos, D., Maoleekoonpiroj, S., Smylie, M., Martins, R., et al. (2005). Erlotinib in previously treated non-small-cell lung cancer. N. Engl. J. Med. *353*, 123–132.

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. Predict. Funct. Mol. Phenotypic Consequences Amin. Acid Substitutions Using Hidden Markov Model. *34*, 57–65.

Shirley, M. (2019). Rucaparib: A Review in Ovarian Cancer. Target. Oncol.

Shlien, A., and Malkin, D. (2009). Copy number variations and cancer. Genome Med.

Sidhu, S.S., Fairbrother, W.J., and Deshayes, K. (2003). Exploring protein-protein interactions with phage display. ChemBioChem.

da Silva, J.P.M., Acencio, M.L., Mombach, J.C.M., Vieira, R., da Silva, J.C., Lemke, N., and Sinigaglia, M. (2008). In silico network topology-based prediction of gene essentiality. Phys. A Stat. Mech. Its Appl. *387*, 1049–1055.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W452–W457.

Sismani, C., Koufaris, C., and Voskarides, K. (2015). Copy number variation in human health, disease and evolution. In Genomic Elements in Health, Disease and Evolution: Junk DNA, p.

Slamon, D., Clark, G., Wong, S., Levin, W., Ullrich, A., and McGuire, W. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science (80-.).

Söderberg, O., Leuchowius, K.J., Gullberg, M., Jarvius, M., Weibrecht, I., Larsson, L.G., and Landegren, U. (2008). Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. Methods.

Srihari, S., Singla, J., Wong, L., and Ragan, M.A. (2015). Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. Biol. Direct.

Stark, C. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*, D535–D539.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400–404.

Stinchcombe, T.E., and Socinski, M.A. (2008). Gefitinib in advanced non-small cell lung cancer: does it deserve a second chance? Oncologist *13*, 933–944.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., and Inouye, M. (1966). Frameshift Mutations and the Genetic Code. Cold Spring Harb. Symp. Quant. Biol.

Strobl, C., Boulesteix, A.L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics.

Syed, Y.Y. (2017). Rucaparib: First Global Approval. Drugs.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M.S., Getz, G., Bader, G.D., Ding, L., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci. Rep. *3*, 2650.

Tangutoori, S., Baldwin, P., and Sridhar, S. (2015). PARP inhibitors: A new era of targeted therapy. Maturitas *81*, 5–9.

Tarca, A.L., Carey, V.J., Chen, X. wen, Romero, R., and Drăghici, S. (2007). Machine learning and its applications to biology. PLoS Comput. Biol.

Thatcher, N., Chang, A., Parikh, P., Rodrigues Pereira, J., Ciuleanu, T., von Pawel, J., Thongprasert, S., Tan, E.H., Pemberton, K., Archer, V., et al. (2005). Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer). Lancet *366*, 1527–1537.

Theofilatos, K., Dimitrakopoulos, C., Likothanassis, S., Kleftogiannis, D., Moschopoulos, C., Alexakos, C., Papadimitriou, S., and Mavroudi, S. (2014). The Human Interactome Knowledge Base (HINT-KB): an integrative human protein interaction database enriched with predicted protein–protein interaction scores using a novel hybrid technique. Artif. Intell. Rev. *42*, 427–443.

Thoma, C.R., Toso, A., Gutbrodt, K.L., Reggi, S.P., Frew, I.J., Schraml, P., Hergovich, A., Moch, H., Meraldi, P., and Krek, W. (2009). VHL loss causes spindle misorientation and chromosome instability. Nat. Cell Biol. *11*, 994–1001.

Thompson, S.L., and Compton, D.A. (2010). Proliferation of aneuploid human cells is limited by a p53-dependent mechanism. J. Cell Biol. *188*, 369–381.

Tlsty, T.D., Briot, A., Gualberto, A., Hall, I., Hess, S., Hixon, M., Kuppuswamy, D., Romanov, S., Sage, M., and White, A. (1995). Genomic instability and cancer. Mutat. Res. Repair.

Tomasini, R., Mak, T.W., and Melino, G. (2008). The impact of p53 and p73 on aneuploidy and cancer. Trends Cell Biol. *18*, 244–252.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. *19*, A68-77.

Tong, A.H.Y., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C.W. V, Bussey, H., et al. (2001a). Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science (80-.). *294*, 2364–2368.

Tong, A.H.Y., 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 (80-.). *303*, 808–813.

Tong, A.H.Y.Y., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C.W.V. V, Bussey, H., et al. (2001b). Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science (80-.). *294*, 2364–2368.

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.e16.

Tsugane, S. (2005). Salt, salted food intake, and risk of gastric cancer: Epidemiologic evidence. Cancer Sci.

Turajlic, S., Larkin, J., and Swanton, C. (2015). SnapShot: Renal Cell Carcinoma. Cell *163*, 1556-1556e1.

Turner, N.C., Lord, C.J., Iorns, E., Brough, R., Swift, S., Elliott, R., Rayter, S., Tutt, A.N., and Ashworth, A. (2008). A synthetic lethal siRNA screen identifying genes mediating sensitivity to a PARP inhibitor. EMBO J. *27*, 1368–1377.

Ulitsky, I., and Shamir, R. (2007). Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. Mol. Syst. Biol. *3*.

Usaj, M., Tan, Y., Wang, W., VanderSluis, B., Zou, A., Myers, C.L., Costanzo, M., Andrews, B., and Boone, C. (2017). TheCellMap.org: A Web-Accessible Database for Visualizing and Mining the Global Yeast Genetic Interaction Network. G3 (Bethesda). *7*, 1539–1549.

Valsesia, A., Macé, A., Jacquemont, S., Beckmann, J.S., and Kutalik, Z. (2013). The growing importance of CNVs: New insights for detection and clinical interpretation. Front. Genet.

Varmus, H., and Kumar, H.S. (2013). Addressing the Growing International Challenge of Cancer: A Multinational Perspective. Sci. Transl. Med. *5*, 175cm2-175cm2.

Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nat. Protoc.

Wang, X., and Simon, R. (2013). Identification of potential synthetic lethal genes to p53 using a computational biology approach. BMC Med. Genomics *6*, 30.

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 (80-.). *350*, 1096–1101.

Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., and Bryant, S.H. (2009). PubChem: A public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. *37*.

Weiderpass, E. (2010). Lifestyle and Cancer Risk. J. Prev. Med. Public Heal. *43*, 459.

Weinstein, I.B. (2002). Cancer: Addiction to oncogenes - The Achilles heal of cancer. Science (80-.).

Weisberg, E., Manley, P.W., Cowan-Jacob, S.W., Hochhaus, A., and Griffin, J.D. (2007). Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukaemia. Nat. Rev. Cancer.

Welsh MJ, Ramsey BW, Accurso F, C.G. (2001). OMMBID | Content. In The Metabolic and Molecular Basis of Inherited Diseases, pp. 5121–5188.

Wen, J., and Brogna, S. (2008). Nonsense-mediated mRNA decay. Biochem. Soc. Trans.

Williamson, C.T., Miller, R., Pemberton, H.N., Jones, S.E., Campbell, J., Konde, A., Badham, N., Rafiq, R., Brough, R., Gulati, A., et al. (2016). ATR inhibitors as a synthetic lethal therapy for tumours deficient in ARID1A. Nat. Commun. *7*, 13837.

Wishart, D.S. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. *34*, D668–D672.

Witsch, E., Sela, M., and Yarden, Y. (2010). Roles for Growth Factors in Cancer Progression. Physiology *25*, 85–101.

Wong, S.L., Zhang, L. V, Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., et al. (2004a). Combining biological networks to predict genetic interactions. Proc. Natl. Acad. Sci. U. S. A. *101*, 15682–15687.

Wong, S.L., Zhang, L. V., Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., et al. (2004b). Combining biological networks to predict genetic interactions. Proc. Natl. Acad. Sci. U. S. A. *101*, 15682–15687.

Woods, D., and Turchi, J.J. (2013). Chemotherapy induced DNA damage response Convergence of drugs and pathways. Cancer Biol. Ther.

Workman, P., Al-Lazikani, B., and Clarke, P.A. (2013). Genome-based cancer therapeutics: targets, kinase drug resistance and future strategies for precision oncology. Curr. Opin. Pharmacol. *13*, 486–496.

Wu, M., Li, X.X., Zhang, F., Li, X.X., Kwoh, C., and Zheng, J. (2013). Meta-analysis of Genomic and Proteomic Features to Predict Synthetic Lethality of Yeast and Human Cancer. Proc. Int. Conf. Bioinformatics, Comput. Biol. Biomed. Informatics 384–391.

Wu, M., Li, X.X.X., Zhang, F., Kwoh, C.-K., and Zheng, J. (2014). In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. Cancer Inform. *13*, 71–80.

Xenarios, I. (2000). DIP: the Database of Interacting Proteins. Nucleic Acids Res.

Yamaoka, K. (2012). Random Forest. J. Inst. Image Inf. Telev. Eng.

Yap, T.A., and Workman, P. (2012). Exploiting the Cancer Genome: Strategies for the Discovery and Clinical Development of Targeted Molecular Therapeutics. Annu. Rev. Pharmacol. Toxicol. *52*, 549–573.

Yarbro, J.W. (1992). Oncogenes and cancer suppressor genes. Semin. Oncol. Nurs. *8*, 30–39.

Yarden, Y., and Ullrich, A. (1988). Molecular Analysis of Signal Transduction by Growth Factors. Biochemistry.

You, Z.H., Yin, Z., Han, K., Huang, D.S., and Zhou, X. (2010). A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. BMC Bioinformatics *11*, 343.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. Bioinformatics *26*, 976–978.

Yu, M.K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J.F., Ng, C.T., Krogan, N., Sharan, R., and Ideker, T. (2016). Translation of genotype to phenotype by a hierarchy of cell subsystems. Cell Syst. *2*, 77–88.

Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.-Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

Zaini, Z.M., McParland, H., Møller, H., Husband, K., and Odell, E.W. (2018). Predicting malignant progression in clinically high-risk lesions by DNA ploidy analysis and dysplasia grading. Sci. Rep.

Zhang, Y.Q., and Rajapakse, J.C. (2008). Machine Learning in Bioinformatics.

Zhang, F., Fan, Z., Min, W., Xue-Juan, L., Xiao-Li, L., Kwoh, C.K., Jie, Z., Wu, M., Li, X.-J.X.-L., Li, X.-J.X.-L., et al. (2015). Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. J. Bioinform. Comput. Biol. *13*, 1541002.

Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., et al. (2011). International cancer genome consortium data portal-a one-stop shop for cancer genomics data. Database *2011*, bar026.

Zhang, X., Acencio, M.L., and Lemke, N. (2016). Predicting essential genes and proteins based on machine learning and network topological features: A comprehensive review. Front. Physiol. *7*, 1–11.

Zhong, W., and Sternberg, P.W. (2006). Genome-wide prediction of C. elegans genetic interactions. Science (80-.). *311*, 1481–1484.

9 - Appendices

9.1 Appendix 1 – Contribution to other work

Mutational patterns in oncogenes and tumour suppressors

https://doi.org/10.1042/BST20160001

Hanadi M. Baeissa, Graeme Benstead-Hume, Christopher J. Richardson, Frances M.G. Pearl

June 2016

Abstract

All cancers depend upon mutations in critical genes, which confer a selective advantage to the tumour cell. Knowledge of these mutations is crucial to understanding the biology of cancer initiation and progression, and to the development of targeted therapeutic strategies. The key to understanding the contribution of a disease-associated mutation to the development and progression of cancer, comes from an understanding of the consequences of that mutation on the function of the affected protein, and the impact on the pathways in which that protein is involved. In this paper we examine the mutation patterns observed in oncogenes and tumour suppressors, and discuss different approaches that have been developed to identify driver mutations within cancers that

contribute to the disease progress. We also discuss the MOKCa database where we have developed an automatic pipeline that structurally and functionally annotates all proteins from the human proteome that are mutated in cancer.

Contribution

For this paper I provided informatics and statistical analysis support.

Identification and analysis of mutational hotspots in oncogenes and tumour suppressors

https://doi.org/10.18632/oncotarget.15514

Hanadi Baeissa, Graeme Benstead-Hume, Christopher J. Richardson, Frances M.G. Pearl

February 2017

Abstract

The key to interpreting the contribution of a disease-associated mutation in the development and progression of cancer is an understanding of the consequences of that mutation both on the function of the affected protein and on the pathways in which that protein is involved. Protein domains encapsulate function and position-specific domain based analysis of mutations have been shown to help elucidate their phenotypes.

In this paper we examine the domain biases in oncogenes and tumour suppressors, and find that their domain compositions substantially differ. Using data from over 30 different

cancers from whole-exome sequencing cancer genomic projects we mapped over one million mutations to their respective Pfam domains to identify which domains are enriched in any of three different classes of mutation; missense, indels or truncations. Next, we identified the mutational hotspots within domain families by mapping small mutations to equivalent positions in multiple sequence alignments of protein domains

We find that gain of function mutations from oncogenes and loss of function mutations from tumour suppressors are normally found in different domain families and when observed in the same domain families, hotspot mutations are located at different positions within the multiple sequence alignment of the domain.

By considering hotspots in tumour suppressors and oncogenes independently, we find that there are different specific positions within domain families that are particularly suited to accommodate either a loss or a gain of function mutation. The position is also dependent on the class of mutation.

We find rare mutations co-located with well-known functional mutation hotspots, in members of homologous domain superfamilies, and we detect novel mutation hotspots in domain families previously unconnected with cancer. The results of this analysis can be accessed through the MOKCa database (http://strubiol.icr.ac.uk/extra/MOKCa).

Contribution

I coded a binomial test to identify which positions had a significant number of mutations. If each individual mutation were to affect a random residue across the domain the frequency of mutations at each site would follow a binomial distribution. As such our null model states that there is an equal probability of a mutation occurring at each residue on the given domain.

Where n is the total number of mutations in the domain, k is the number of mutations falling at a specific residue and p the probability of any mutation affecting a specific residue we can find the probability of observing k mutations falling at any specific point in the domain by calculating the probability of a minimum of k mutations at that point and comparing it to our null model.

‘Big data’ approaches for novel anti-cancer drug discovery.

https://doi.org/10.1080/17460441.2017.1319356

Benstead-Hume G., Wooller SK, Pearl FMG. Expert Opinion on Drug Discovery 12(6):1-11

May 2017

Abstract

The development of improved cancer therapies is frequently cited as an urgent unmet medical need. Here we review how recent advances in platform technologies and the increasing availability of biological ‘big data’ are providing an unparalleled opportunity to systematically identify the key genes and pathways involved in tumorigenesis. We then discuss how these discoveries may be amenable to therapeutic interventions.

We discuss the current approaches that use ‘big data’ to identify cancer drivers. These approaches include genomic sequencing, pathway data, multi-platform data, identifying genetic interactions such as synthetic lethality and using cell line data. We review how

big data is being used to assess the tractability of potential drug targets and how systems biology is being utilised to identify novel drug targets. We finish the review with an overview of available data repositories and tools being used at the forefront of cancer drug discovery. Targeted therapies based on the genomic events driving the tumour will eventually inform treatment protocols. However, using a tailored approach to treat all tumour patients may require developing a large repertoire of targeted drugs.

Contribution

I researched and wrote this paper with S.K.W. and F.M.G.P.

Bioinformatics in translational drug discovery

http://doi.org/10.1042/BSR20160180

Sarah K. Wooller, Graeme Benstead-Hume, Xiangrong Chen, Yusuf Ali, Frances M.G. Pearl

July 2017

Abstract

Bioinformatics approaches are becoming ever more essential in translational drug discovery both in academia and within the pharmaceutical industry. Computational exploitation of the increasing volumes of data generated during all phases of drug discovery is enabling key challenges of the process to be addressed. Here, we highlight some of the areas in which bioinformatics resources and methods are being developed to support the drug discovery pipeline. These include the creation of large data warehouses, bioinformatics algorithms to analyse 'big data' that identify novel drug targets and/or biomarkers, programs to assess the tractability of targets, and prediction

of repositioning opportunities that use licensed drugs to treat additional indications.

Contribution

I researched and wrote this paper with S.K.W. and F.M.G.P.

Repression of Transcription at DNA Breaks Requires Cohesin throughout Interphase and Prevents Genome Instability

https://doi.org/10.1016/j.molcel.2018.11.001

Cornelia Meisenberg, Sarah I.Pinder, Suzanna R.Hopkins, Sarah K.Wooller, GraemeBenstead-Hume, Frances M.G.Pearl, Penny A.Jeggo, Jessica A.Downs

13 December 2018

Abstract

Cohesin subunits are frequently mutated in cancer, but how they function as tumor suppressors is unknown. Cohesin mediates sister chromatid cohesion, but this is not always perturbed in cancer cells. Here, we identify a previously unknown role for cohesin. We find that cohesin is required to repress transcription at DNA double-strand breaks (DSBs). Notably, cohesin represses transcription at DSBs throughout interphase, indicating that this is distinct from its known role in mediating DNA repair through sister chromatid cohesion. We identified a cancer-associated SA2 mutation that supports sister chromatid cohesion but is unable to repress transcription at DSBs. We further show that failure to repress transcription at DSBs leads to large-scale genome rearrangements. Cancer samples lacking SA2 display mutational patterns consistent with loss of this pathway. These findings uncover a new function for cohesin that provides insights into its

frequent loss in cancer.

Contribution

To investigate the difference in mutational patterns in SA2 competent and SA2 deficient tumors, I implemented mutational fingerprints for two groups of patients were generated using mutational data from whole genome screens annotated in the COSMIC database. One group included samples from 336 bladder cancer patients that did not exhibit a SA2 mutation and the other group included samples from 38 bladder cancer patients with a SA2 mutation.

For both groups of samples their mutational fingerprints were decomposed using a non-negative matrix factorisation to produce 5 signatures. Decomposition was performed using the Brunet method through the NMF library in R3.4.0. The resulting signatures were compared to those published in COSMIC using a correlation matrix produced again in R using the Pearson's correlation method.

Cell-derived extracellular vesicles can be used as a biomarker reservoir for glioblastoma tumor subtyping

https://doi.org/10.1038/s42003-019-0560-x

Rosemary Lane, Thomas Simon, Marian Vintu, Benjamin Solkin, Barbara Koch, Nicolas Stewart, Graeme Benstead-Hume, Frances M. G. Pearl, Giles Critchley, Justin Stebbing & Georgios Giamas

August 2019

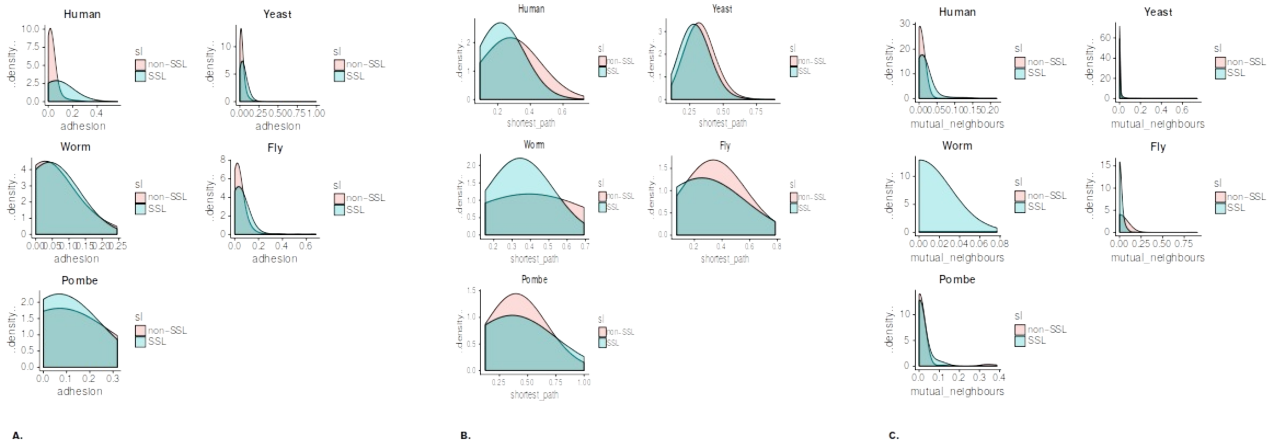
Glioblastoma (GBM) is one of the most aggressive solid tumors for which treatment options and biomarkers are limited. Small extracellular vesicles (sEVs) produced by both GBM and stromal cells are central in the inter-cellular communication that is taking place in the tumor bulk. As tumor sEVs are accessible in biofluids, recent reports have suggested that sEVs contain valuable biomarkers for GBM patient diagnosis and follow-up. The aim of the current study was to describe the protein content of sEVs produced by different GBM cell lines and patient-derived stem cells. Our results reveal that the content of the sEVs mirrors the phenotypic signature of the respective GBM cells, leading to the description of potential informative sEV-associated biomarkers for GBM subtyping, such as CD44. Overall, these data could assist future GBM in vitro studies and provide insights for the development of new diagnostic and therapeutic methods as well as personalized treatment strategies.

Contribution

Mean phenotype parameter measurement across all available cell lines was decomposed into seven different signatures to reduce the dimensionality of the data and provide a method of clustering the cell lines by phenotype similarity. Additionally the LN18, U87, U118, G166, and GS090 cell lines alone were decomposed into four signatures. This decomposition was achieved using non-negative matrix factorization (NMF). Each cell-line’s mean parameter measurement was used to build a feature matrix. NMF was used to decompose these features into two separate matrices, the basis, which describes the composition of each signature and the coefficient, which reports how prominent each signature is in each cell line and stem cell. The number of components parameter used for each decomposition was decided by running many NMF decompositions with increasing parameters, and choosing the number of components where the reconstruction error plateaued. Finally, we used hierarchical clustering on the coefficient matrices in order to cluster GBM cell lines and stem cells based on signature composition similarity.

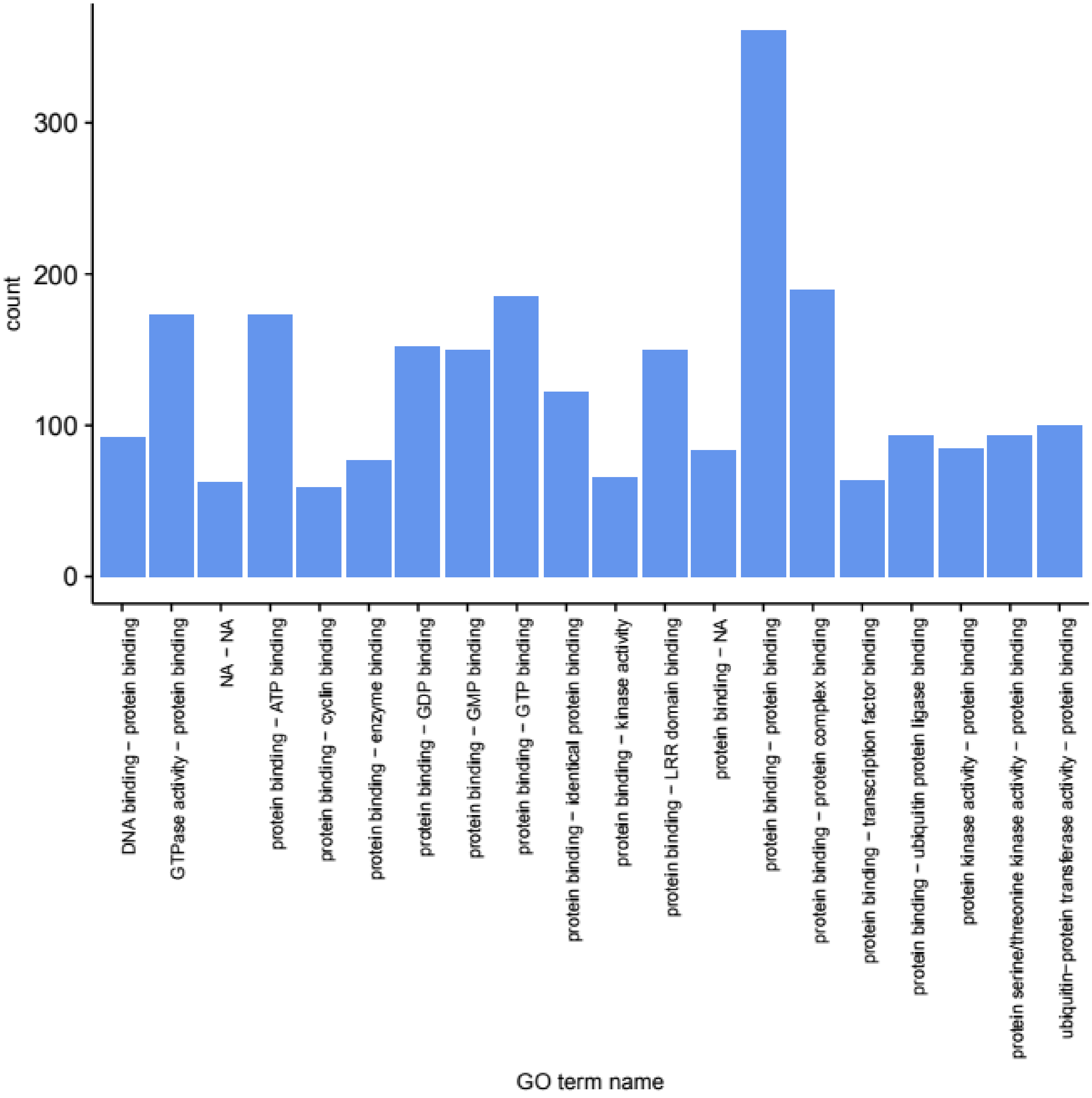
9.2 Appendix 2 - Chapter 3 supplementary content

9.2.1 A2.1 Chapter 3 supplementary figures

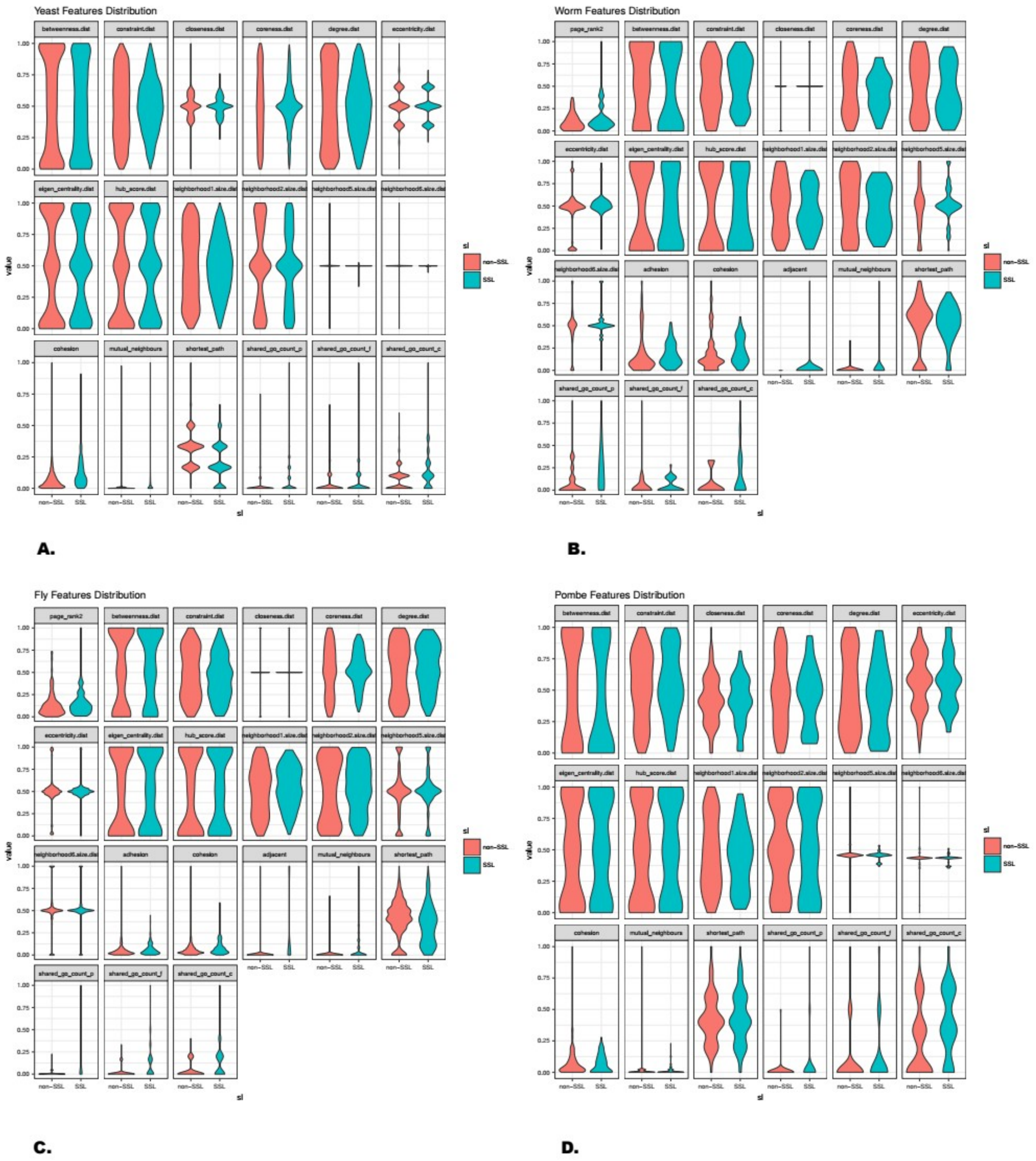


Supplementary Figure 3.1

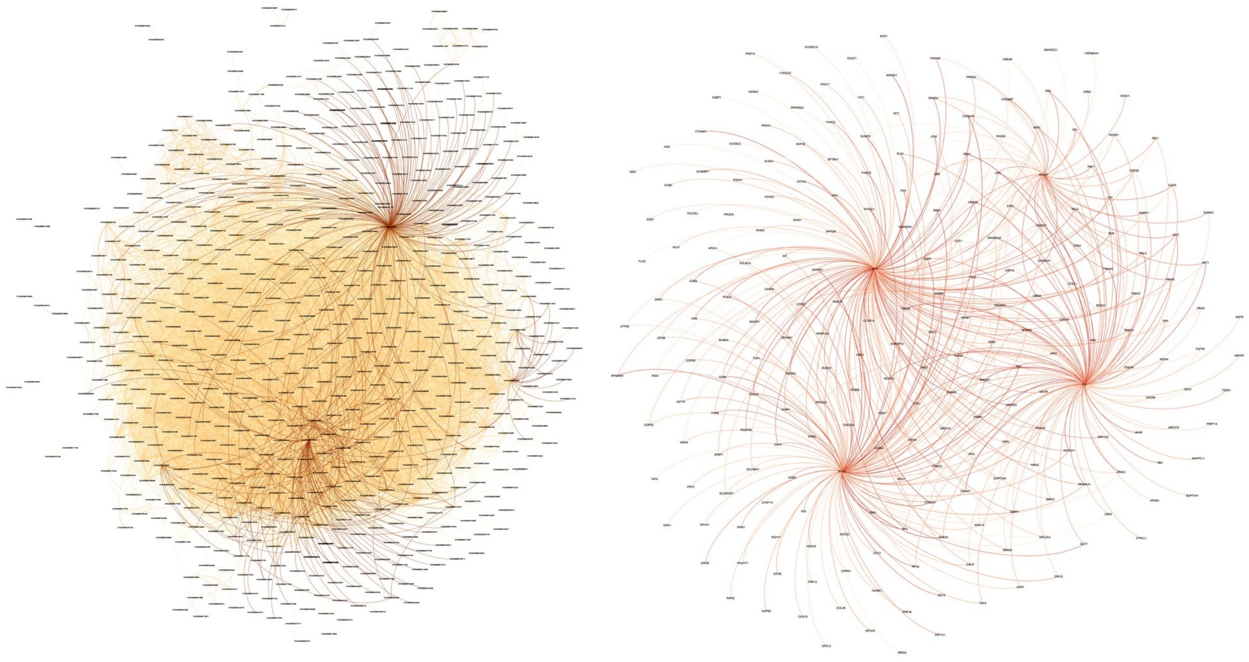
a. A distribution of normalised adhesion scores for each organism illustrate significant differences in SSL and non-SSL pairs across species. b. A normalised shortest path distribution shows a general trend for shorter shortest paths between H. sapiens SSL pairs though this difference is less pronounced in our model organisms. c. A distribution of normalised mutual neighbour counts suggests that SSL pairs often share more mutual neighbours than non-SSL pairs.



Supplementary Figure 3.2. Count of most common associations between molecular function GO terms observed in SSL pairs. Individual feature GO associations extracted from full GO annotation lists for each SSL gene pair.



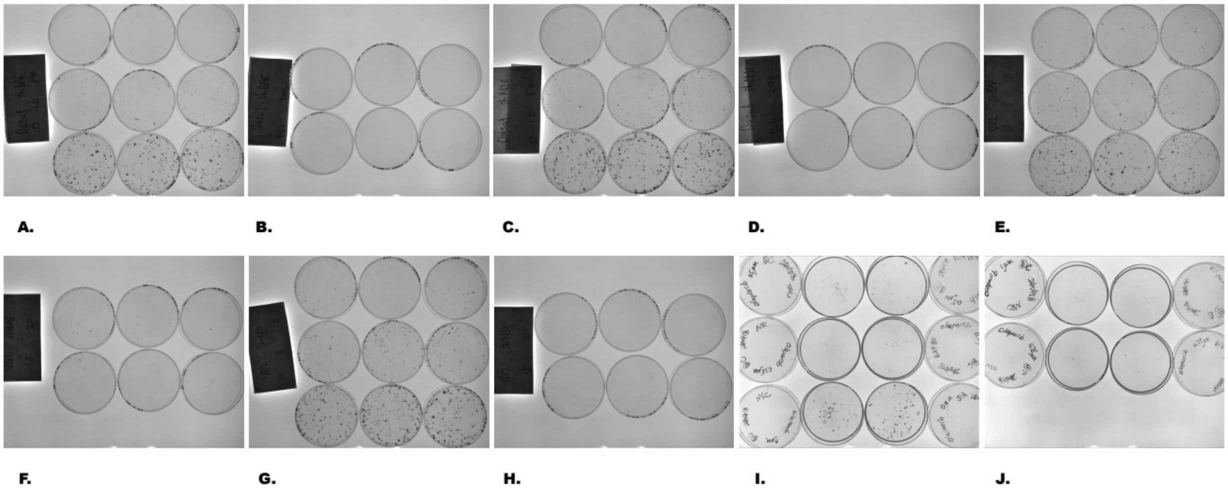
Supplementary Figure 3.3. Violin plots illustrating feature value distributions for a. S. cerevisiae, b. C. elegans, c. D. melanogaster and d. S. pombe.



Supplementary Figure 3.4

a. *Full SSL interaction network of predicted human SSL pairs shaded by likelihood of being a true SSL pair based on consensus score. Red edges are interactions sourced from our training data (directly from BioGRID) lighter edges denote a lower consensus scores. Produced with Gephi 0.9.1 (Bastian et al., 2009).*

b. *Network of SSL interaction predictions with high consensus scores associated with known tumour suppressors including, where available, VHL, BRCA1, BRCA2, PBRM1, PTEN and APC.*



Supplementary Figure 3.5. Carcinogenic survival assay plate images for ABL inhibitor Dasatanib (marked as Dasat) (supp figures 5a, b, c & d) and POLA inhibitor Erocalciferol (marked as VD2, an abbreviation of vitamin D2) experiments (supp figures 5e, f, g & h). BAF180 knock-out cell-line plate images for the PARP1 inhibitor Olaparib BAF180 are labeled with BAF and control plates marked with NSC on plate lids and the corresponding plate colonies are displayed adjacent to each lid (supp figures 5i & j).

9.2.2 A2.2 Chapter 3 Supplementary Tables

GO Annotation type	***H. sapiens* shared GO terms**	**SSL**	**Non-SSL**	**Welch 2 sample t-test (p)**
Molecular function	Mean	1.195	0.733	**< $2.2*10^{-16}$**
	Stdev.	0.903	0.682	
Biological process	Mean	0.837	0.149	**< $2.2*10^{-16}$**
	Stdev.	1.554	0.550	
Cellular compartment	Mean	1.679	0.817	**< $2.2*10^{-16}$**
	Stdev.	1.210	0.991	

Supplementary Table 3.1a

GO Annotation type	***S. cerevisiae* shared GO terms**	**SSL**	**Non-SSL**	**Welch 2 sample t-test (p)**
Molecular function	Mean	0.486	0.258	**< $2.2*10^{-16}$**
	Stdev.	1.099	0.681	
Biological process	Mean	0.760	0.136	**< $2.2*10^{-16}$**
	Stdev.	1.364	0.504	
Cellular	Mean	1.355	0.674	**< $2.2*10^{-16}$**

compartment	Stdev.	1.369	0.825	

Supplementary Table 3.1b

GO Annotation type	***C. elegans* shared GO terms**	**SSL**	**Non-SSL**	**Welch 2 sample t-test (p)**
Molecular function	Mean	0.423	0.229	**$< 2.2*10^{-16}$**
	Stdev.	0.970	0.643	
Biological process	Mean	2.368	0.683	**$2.642*10^{-07}$**
	Stdev.	2.290	1.210	
Cellular compartment	Mean	0.483	0.256	**$1.89*10^{-08}$**
	Stdev.	0.772	0.510	

Supplementary Table 3.1c

GO Annotation type	***D. melanogaster* shared GO terms**	**SSL**	**Non-SSL**	**Welch 2 sample t-test (p)**
Molecular function	Mean	0.554	0.099	**$< 2.2*10^{-16}$**
	Stdev.	0.843	0.332	
Biological process	Mean	1.842	0.098	**$< 2.2*10^{-16}$**
	Stdev.	2.993	0.332	

Cellular compartment	Mean	0.619	0.179	**< 2.2*10^{-16}**
	Stdev.	0.829	0.428	

Supplementary Table 3.1d

GO Annotation type	***S. pombe*** **shared GO terms**	**SSL**	**Non-SSL**	**Welch 2 sample t-test (p)**
Molecular function	Mean	0.252	0.107	**0.2934**
	Stdev.	0.519	0.331	
Biological process	Mean	0.0841	0.025	**2.832*10^{-09}**
	Stdev.	0.312	0.173	
Cellular compartment	Mean	1.16	0.623	**< 2.2*10^{-16}**
	Stdev.	0.837	0.776	

Supplementary Table 3.1e

Supplementary Table 3.1. Distribution of shared molecular function, biological process and cellular compartment GO terms that occur between SSL and non-SSL pairs in ***a.*** *H. sapiens****, b.*** *S. cerevisea,* ***c.*** *C. elegans,* ***d.*** *D. melanogaster, and* ***e.*** *S. pombe. We observe that SSL pairs share significantly more molecular function and cellular compartment GO terms while non-SSL pairs share significantly more biological process terms in H. sapiens. A welch 2 sample t-test was used to measure significance for each annotation. 2.2*10^{-16} was the smallest value availble.*

	H. sapiens	*S. cerevisiae*	*C. elegans*	*D. melanogaster*	*S. pombe*
Adhesion	✔				
Adjacent				✔	
Cohesion	✔	✔			
Mutual neighbours		✔		✔	
Shared GO Count – cellular compartment	✔	✔		✔	✔
Shared GO count – molecular function			✔		
Shared GO Count – biological process	✔		✔	✔	
Coreness	✔	✔	✔		✔
Neighborhood size			✔		✔

Supplementary Table 3.2. A list of most important features for each species reported via the R caret libraries random forest classifier. Feature importance rankings were calculated by measuring the mean decrease in accuracy without each variable across all tree permutations in the random forest.

Model	**Cross-species SDL classification performance (ROC AUC)**

	H. sapiens	*S. cerevisiae*
H. sapiens	0.782	0.754
S. cerevisiae	0.736	0.890
Consensus	0.805	0.918

Supplementary Table 3.3. Cross validation ROC AUC scores for S. cerevisiae and H.sapiens SDL models. The best score for each species model is highlighted in green. Consensus model results are highlighted in blue.

Gene1	**Gene2**	**Consensus score**
SREBF1	VHL	0.810066584
PTEN	SFN	0.808599164
RBX1	VHL	0.808448941
PTEN	CHEK2	0.808017586
UBE2D3	VHL	0.807876266
BRAF	PTEN	0.807264817
FBXW7	VHL	0.806462054
PTEN	CTNND1	0.806458647
GSK3B	VHL	0.805455056
APC	AURKB	0.805347415
APC	CTNND1	0.805335746
PIN1	PTEN	0.805318639
MAP3K1	PTEN	0.805219709
CDKN1B	PTEN	0.804986161
ORC1	PTEN	0.80472497
ARRB2	PTEN	0.804692535
SKP2	PTEN	0.80468009
BUB1B	PTEN	0.803596615
VHL	NOTCH1	0.803412785

Supplementary Table 3.4. Top 20 Predictions featuring common tumour suppressor genes

Drug	**Target**	**Citation**
Olaparib	PARP1	(Tangutoori et al., 2015)
CP-724714	ERBB2/HER2	(Jani et al., 2007)
ZM 336372	RAF1	(Hall-Jackson et al., 1999)
Ergocalciferol (Vitamin D2)	POLA1	(Mizushina et al., 2003)
AZD1480	JAK2	(Plimack et al., 2013)
Dasatinib	ABL1	(Raju et al., 2012)
CHIR-98014	GSK3B	(Rahmani et al., 2013)

Supplementary Table 3.5. We chose a group of genes with selective inhibitors that were predicted to share a synthetic lethal interaction with BAF180 (PBRM1) for validation. We performed clonogenic survival assays for each inhibitor using U2OS cell lines (shControl + mCherry/NLS and shBAF180 + GFP/NLS).

Organism	**Protein interactions**	**SSL / Negative GI count**	**SDL count**
H. sapiens	60,278	411	259
S. cerevisiae	82,480	17,568 (of 395,199)	2,389
C. elegans	36,332	1,237	0
D. melanogaster	34,324	348	0
S. pombe	47,492	3836 (of 35,391)	0

Supplementary Table 3.6. Number of protein-protein interactions used to generate the protein interaction networks for each organism. Number of SSL pairs and SDL pairs sourced for each organism from BioGRID after filtering for distinct pairs that inlcude genes present in the protein interaction network. The SSL pair data for S. cerevisiae were filtered to include only interactions cited in 3 or more papers. SSL pair data for S. pombe were filtered to include only interactions recorded in 2 or more papers.

Feature	**SLant**	**SINaTRA**
Betweenness	x	x
Constraint	x	
Closeness	x	x
Coreness	x	
Degree	x	x
Eccentricity	x	x
Eigen centrality	x	x
Hub score	x	
Neighbourhood n size	x	x
PageRank		x
Adhesion	x	
Cohesion	x	
Communicability		x
Current-flow betweenness centrailty		x
Adjacent	x	

Mutual neighbours	x	x
Mutual non-neighbours		x
Shortest path	x	
Inversed shortest path		x
Between community	x	
Cross community	x	
Shared GO count – Biological process	x	
Shared GO count – Molecular function	x	
Shared GO count – Cellular compartment	x	

Supplementary Table 3.7. A comparison of the features used by SLant and SINaTRA. SLant also treats node-wise features differently by providing an averaged difference between node pairs as well as the individual values per gene node.

9.3 Appendix 3 - Chapter 4 Supplementary material

9.3.1 A3.1 Chapter 4 Supplementary Tables

Parameter	Features	Description
Betweenness	Gene1 betweenness Gene 2 betweenness Betweeness difference	Count of shortest paths across the graph that pass through a node.
Constraint	Gene 1 constraint Gene 2 constraint Constraint difference	A measure of how much a node's connections are focused on single cluster of neighbours.
Closeness	Node-wise	The number of steps required to reach all other nodes from a given node.
Coreness	Node-wise	Whether a node is part of the k-core of the full graph, the k-core being a maximal sub-graph in which each node has at least degree k.
Degree	Node-wise	The number of edges coming in to or out of the node.
Eccentricity	Node-wise	The shortest path distance from the node farthest from the given node.
Eigen centrality	Node-wise	A measure of how well connected a given node is to other well-connected nodes.
Hub score	Node-wise	Related to the concepts of hubs and authorities the hub score is a measure of how many well linked hubs the

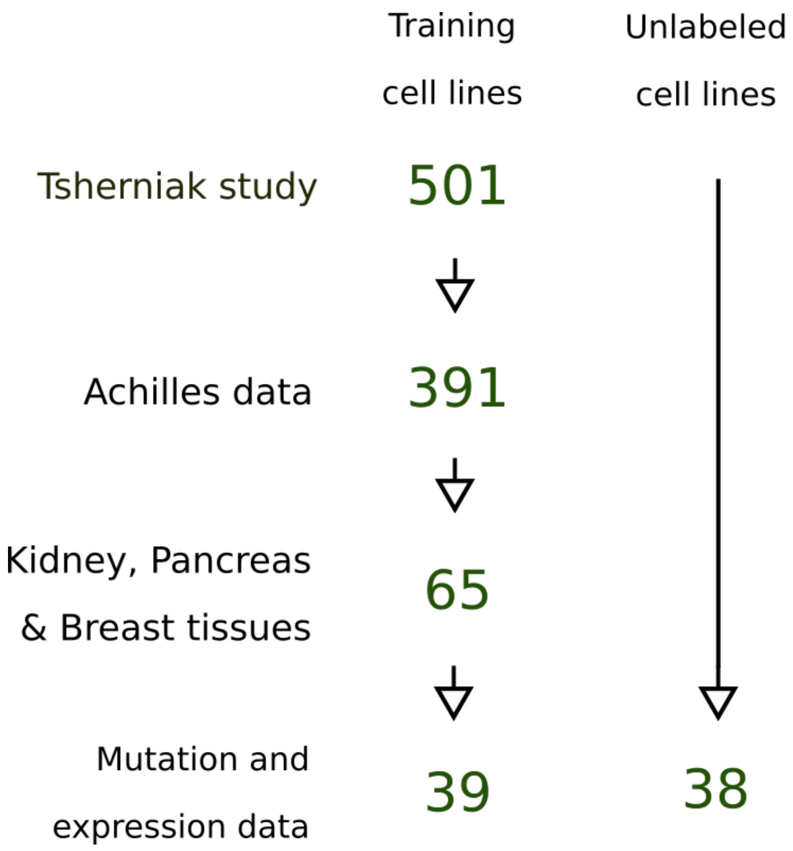
		nodes is linked to.
Neighbourhood n size	Node-wise	The number of nodes within n steps of a given node for n of 1, 2, 5 and 6
Adhesion	Pairwise	The minimum number of edges that would have to be severed to result in two separate sub-graphs separating the source and target nodes.
Cohesion	Pairwise	The minimum number of nodes that would have to be removed to result in two separate sub-graphs separating the source and target nodes.
Adjacent	Pairwise	Whether a source and target node are connected via an edge.
Mutual neighbours	Pairwise	How many first neighbours a target and source node share.
Shortest path	Pairwise	The minimal number of connected vertices that create a path between the source and target node.
Between community	Pairwise	A logical feature stating whether the source and target nodes inhabit the same community produced by the spin glass random walk.
Cross community	Pairwise	A logical feature stating whether the source and target nodes connect two communities as produced by the spin glass random walk.
Shared GO count – Biological process	Go term	The number of biological process GO annotations shared between the source and target node.

Shared GO count – Molecular function	Go term	The number of molecular function GO annotations shared between the source and target node.
Shared GO count – Cellular compartment	Go term	The number of cellular compartment GO annotations shared between the source and target node.

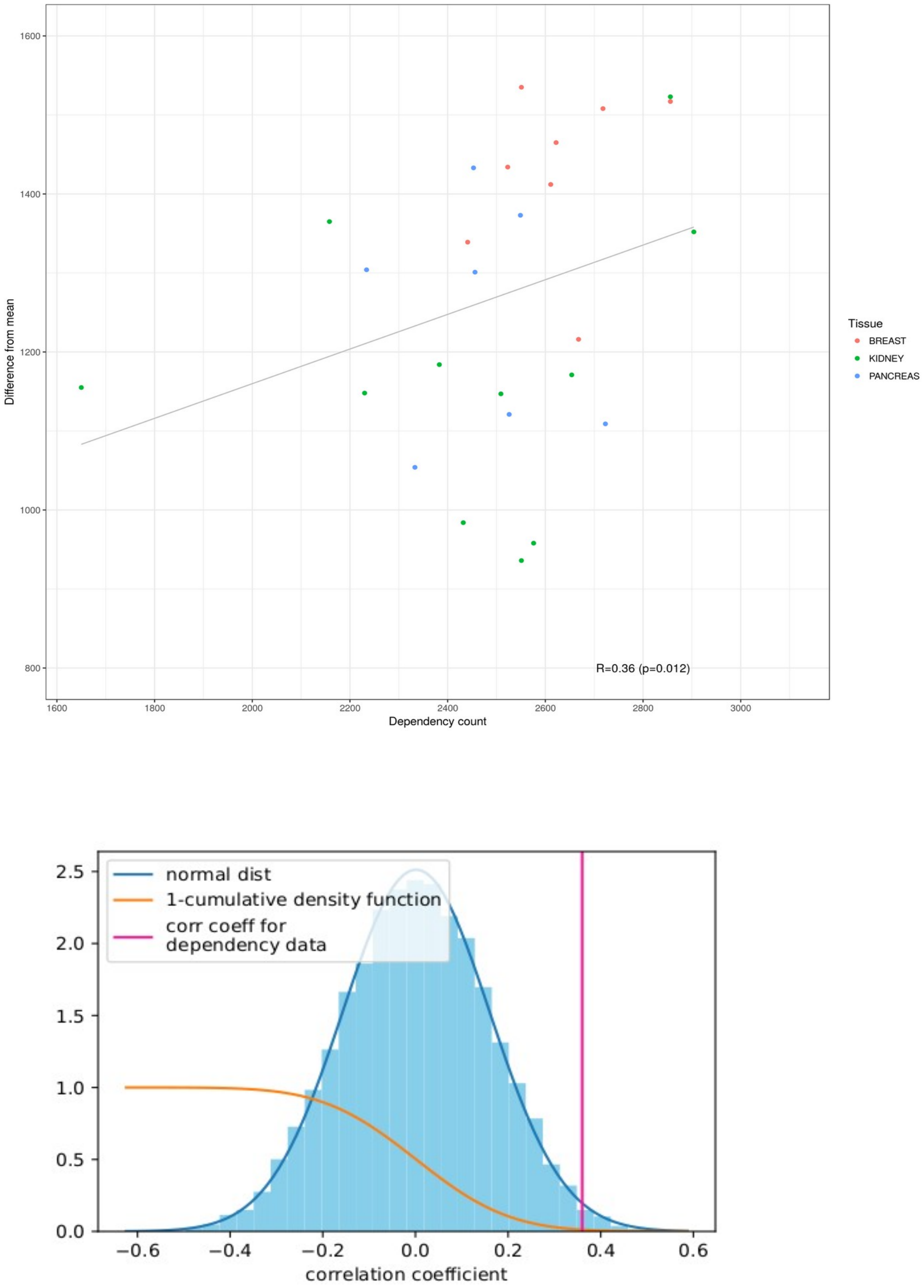
Supplementary Table 4.1 – Features used in the Slant classification models

9.4 Appendix 4 - Chapter 5 Supplementary material

9.4.1 A4.1 Chapter 5 supplementary figures



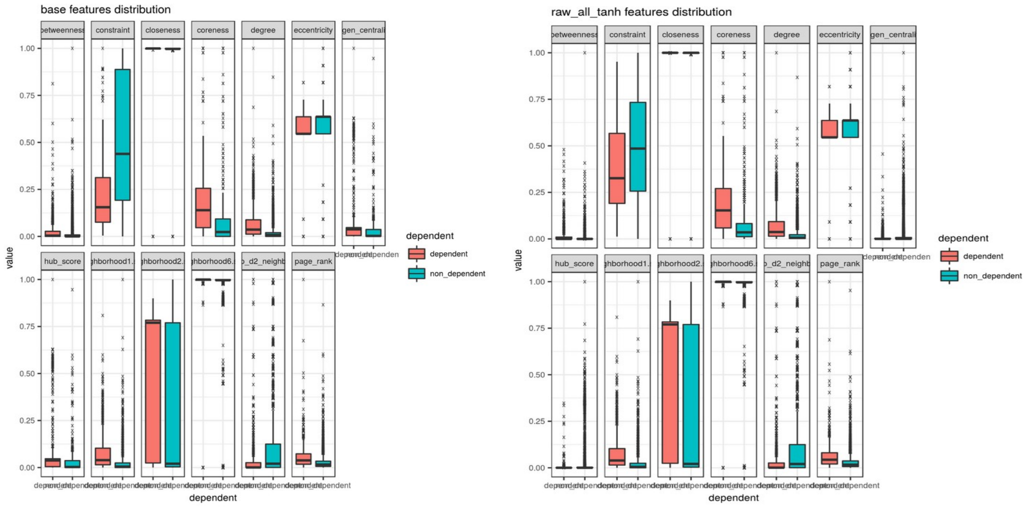
Supplementary Figure 5.1. Number of cell lines available after filtering for public accessibility, tissue type and genetic alteration data availability.



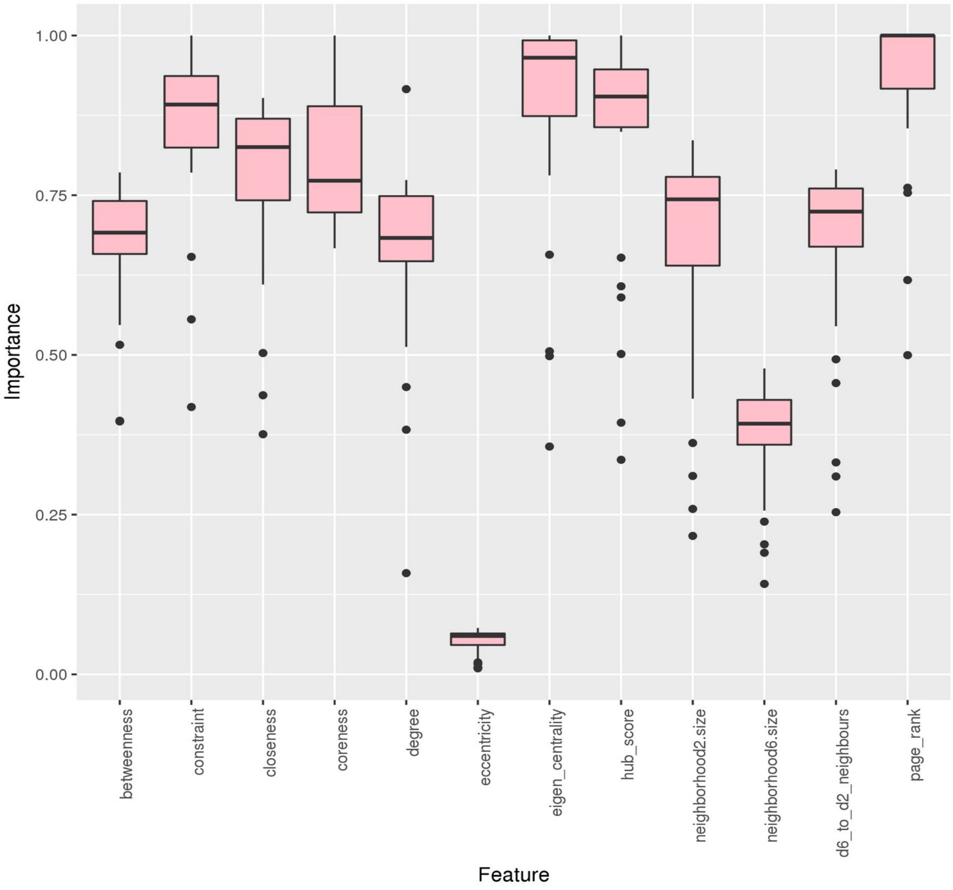
Supplementary Figure 5.2. - Measuring the relationship between generic alteration levels and count of gene dependencies in cell lines

a. By plotting the number of gene dependencies reported for each cell line against a measure of that cell lines genomic alteration we find a small positive correlation between the two.

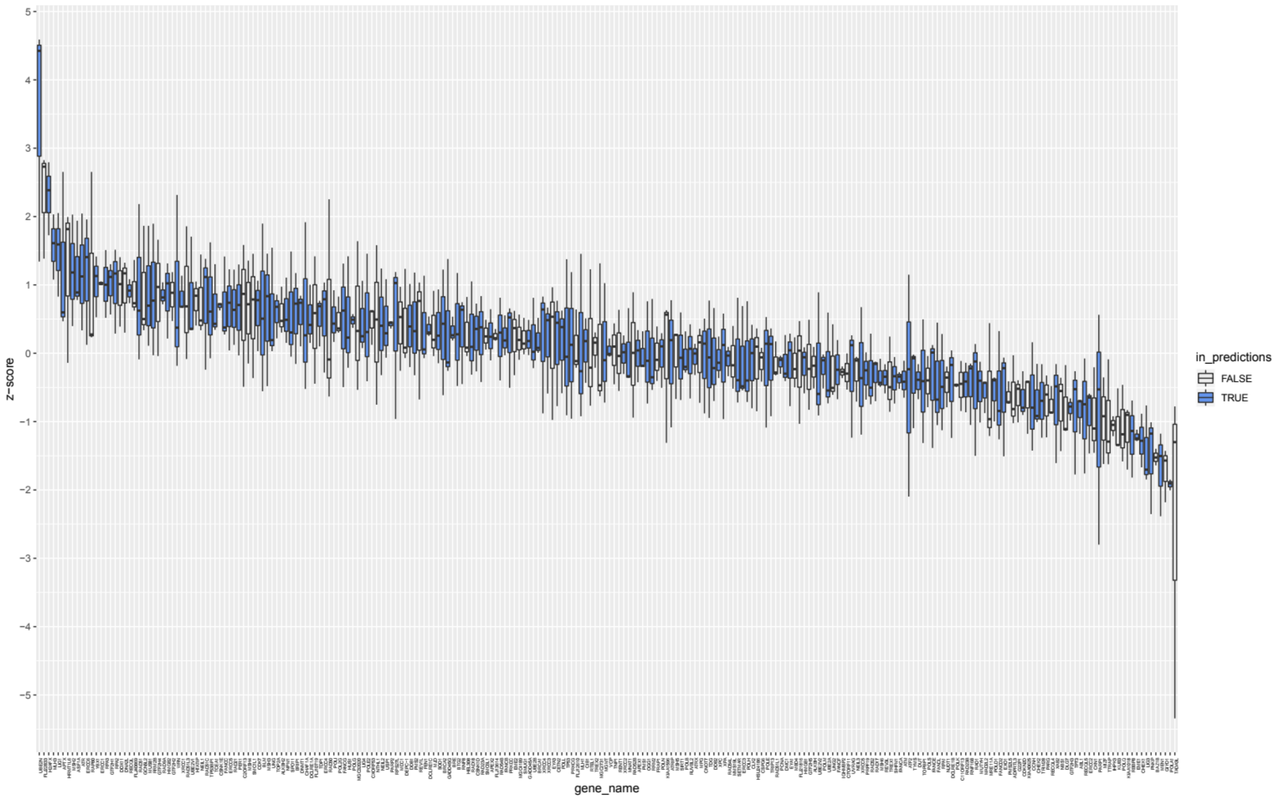
b. Shuffling the data and then finding the correlations for this data demonstrates that our correlation is statistically significant p-value=0.012.



Supplementary Figure 5.3. Feature distributions between dependent and non-dependent gene classes show some   differences between the classes for the betweeness, constraint, eigen centrality and hub_score features.



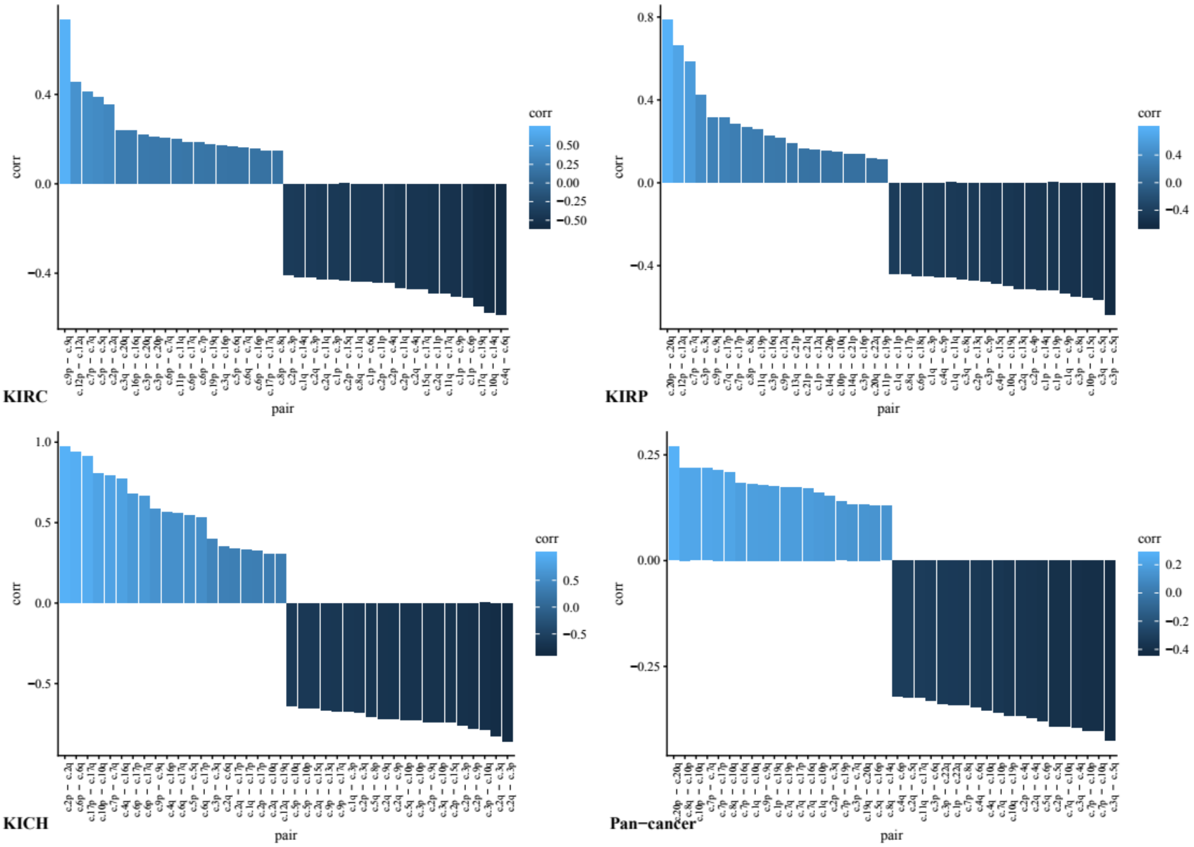
Supplementary Figure 5.4. Importance for each feature used in each model calculated by measuring the mean decrease in accuracy when holding out each variable across all tree permutations in the random forest.



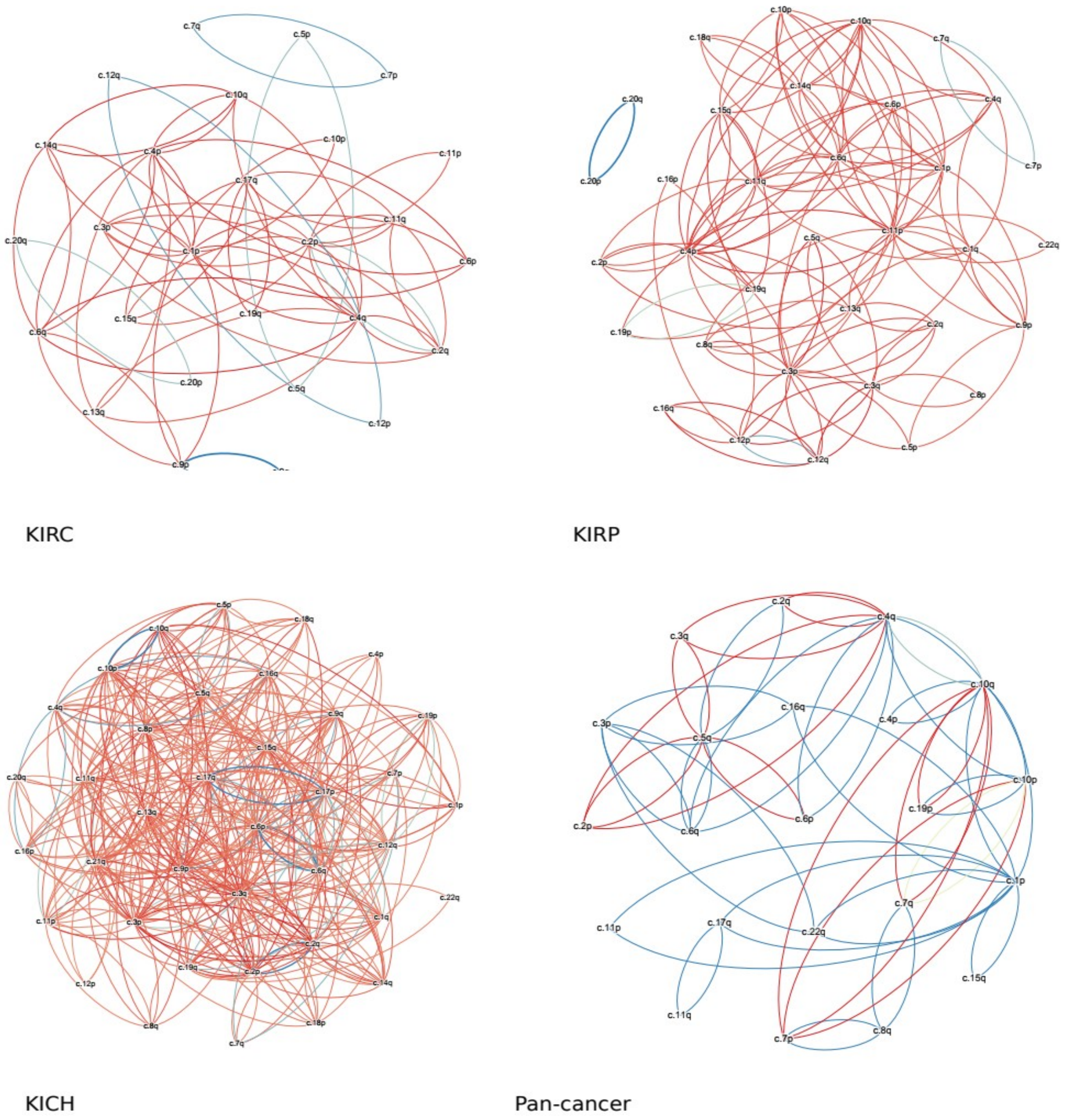
Supplementary Figure 5.5. Survival screen's z-score distribution with variation. This box plot graphs each gene featured in the survival screen with its z-score distribution across 3 experimental repeats. Blue boxes denote genes which were featured in our prediction set. White boxes denote genes that were not in our prediction set due to insufficient training data (i.e. mutational or copy number data).

9.5 Appendix 5 - Chapter 6 Supplementary material

9.5.1 A5.1 Chapter 6 supplementary figures



Supplementary Figure 6.1. A ranked bar chart showing the most extreme arm-wise copy number correlation coefficients.



Supplementary Figure 6.2. A graphical illustration of arm-wise copy number correlation. Nodes represent chromosome arms and edges represent correlations of above a *threshold of r=0.4 and below r=-0.4. Red edges denote positive correlation between the chromosome arms and blue represent negative correlations.*