This dissertation is submitted for the degree of Doctor of Philosophy

Gene expression signatures for cancer cell line drug sensitivity and patient outcome

Michael Schubert

Submitted August 2016

Corpus Christi College,
University of Cambridge
EMBL-EBI

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Michael Schubert
August, 2016

SUMMARY

The explanation of phenotypes in cancer, such as cell line drug response or patient survival, has largely been focussed on genomic alterations. While this approach has generated many profound insights into cancer biology, it does not directly make statements about the signaling impact those cellular aberrations create. With direct measurements much less widely available than gene expression, pathway methods (mostly mapping gene expression onto signalling proteins) have so far fallen short on delivering actionable evidence. This may in part be due to lack of robustness, but these approaches are fundamentally at odds with the notion of tight post-translational control of signal transduction. A way to solve this may be to derive consensus signatures of pathway activity to make inferences about signalling, or signatures of specific drugs to investigate their interactions.

As a baseline (chapter 2), I investigated how well pathway methods compare to driver mutations in terms of explaining cell line drug response. I went on to analyse the value of gene expression as a downstream signature of signaling activity instead of mapping it to the pathway components by means of a previously published platform. This improved over mapping pathway members using Gene Ontology or Reactome, but it considered only sets of up-regulated genes defined by multiple arbitrary cutoffs. Hence, I extended the data set and created a linear model (chapter 3) that compares favourably to gene set as well as state of the art pathway methods in terms of recovering driver mutations and providing biomarkers for cell line drug response and patient survival (chapter 4). To complement this, I investigate how gene expression signatures of drugs can be used in conjunction with viability data to suggest effective drug combinations where no pathway information is available (chapter 5).

To the best of my knowledge, this thesis represents the first comprehensive analysis of different pathway methods across primary cohorts and cancer cell lines, as well as the first large-scale systematic analysis of drug sensitisation that could lead to new drug combinations.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the generous support and feedback of many people, for which I am deeply grateful.

First and foremost, I would like to thank my supervisor Julio Saez-Rodriguez for providing me with the opportunity to undertake my PhD in his group at EMBL-EBI as well as his support during my PhD. I would also like to extend my gratitude to all members of the Saez Rodriguez group, especially Francesco Iorio, Emanuel Goncalves, Luz Garcia Alonso, Michael Menden and Camille Terfve for fruitful dicussions and their friendship.

I would like to thank my academic collaborators, among which are my industry supervisor Joanna Betts, the possibility to visit the US site of GSK with Pankaj Agarwal, the Sanger GDSC team lead by Mathew Garnett and Ultan McDermott, as well as people involved in smaller side projects too numerous to name.

Many thanks to my TAC members Florian Markowetz, Sarah Teichmann, and Christoph Merten who I was always looking forward to meet once a year for a useful critique of my project and much needed direction.

Finally, I would like to thank the EMBL community for an amazing four years that I spent at the institute. You know who you are.

CONTENTS

LIST OF ABBREVIATIONS

ACC	Adrenocortical carcinoma
ALL	Acute Lymphoblastic Leukaemia
ANOVA	Analysis of Variance
ARACNE	Accurate Reconstruction of Cellular Networks
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CCLE	Cancer Cell Line Encyclopedia
ChIP-seq	Chromatin Immunoprecipitation sequencing
COAD	Colon adenocarcinoma
COREAD	Colon and rectum adenocarcinoma
CRAN	Comprehensive R Archive Network
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
GSEA	Gene Set Enrichment Analysis
GC	Gas Chromatography
GDSC	Genomics of Drug Sensitivity in Cancer
GO	Gene Ontology
HGNC	Human Gene Nomenclature Consortium
HNSC	Head and Neck squamous cell carcinoma
HPLC	High Performance Liquid Chromatography
ICGC	International Cancer Genome Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIRC	Kidney renal clear cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LINCS	Library of Integrated Network-Based Cellular Signatures
LPS	Lipopolysaccharide

LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MEME	Multiple EM for Motif Elicitation
MIAME	Minimum Information Required in the Annotation of Microarray Experiments
MINDy	Modulator Inference by Network Dynamics
mRNA	Messenger RNA
NMF	Non-negative Matrix Factorisation
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PARADIGM	PAthway Representation and Analysis by Direct Reference on Graphical Models
PCA	Principal Component Analysis
PolII	DNA-dependent RNA polymerase (Polymerase II)
PRAD	Prostate adenocarcinoma
PRL	Prototype Ranked List
PTM	Post-translational modification
PyPI	Python Packaging Index
READ	Rectum adenocarcinoma
RNA	Ribonucleic Acid
SKCM	Skin Cutaneous Melanoma
SPEED	Signaling Pathway Enrichment using Experimental Data sets
SPIA	Signaling Pathway Impact Analysis
SNP	Single Nucleotide Polymorphism
STAD	Stomach adenocarcinoma
SQLite	Structured Query Language file-based database implementation
SVD	Singular Value Decomposition
TCGA	The Cancer Genome Atlas
THCA	Thyroid carcinoma
TOF	Time of Flight
t-SNE	T-distrubuted Stochastic Neighbor Embedding
UCEC	Uterine Corpus Endometrial Carcinoma
VCS	Version Control System
YAML	Yet Another Markup Language

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Parts of section 1.6 were previously published. The text incorporated represents a draft stage of the article below that was entirely written by myself:

Schubert, M & Iorio, F "*Exploiting combinatorial patterns in cancer genomic data for personalized therapy and new target discovery.*" **Pharmacogenomics** 15, 1943–1946 (2014).

1.1 CANCER BIOLOGY

1.1.1 *Significance and epidemiology*

Cancer is a disease of the genes, given rise by changes in the genome that mediate malignant transformation and hence uncontrolled growth of a cell (Hanahan and Weinberg, 2000). The International Agency for Research on Cancer estimates of global incidence of cancer to be 12.7 million new cases on 7.6 million cancer deaths based on estimates for 182 countries in 2008 (Ferlay et al., 2010). In the United States, it has even surpassed heart disease as the leading cause of death in people younger than 85 (Twombly, 2005). Breast cancer is the most abundant form in females, accounting for a total of 23% of cases and 14% of deaths. For males, the most abundant is lung cancer, comprised of 17% of cases and 23% of cancer-related deaths (Jemal et al., 2011).

Needless to say, the disease is a global health concern and better ways of diagnosis and treatment are needed, as well as a better understanding of the molecular mechanisms.

1.1.2 *Causes of genetic variation*

Germline variation

In the human population, there is natural variation in the genome from one individual to another, which gets passed down through generations in the process of reproduction. Each human being has two copies of their DNA, one passed down from their mother and the other passed down from their father.[1] Disregarding of where the copies that an

1 this is excluding mitochondrial DNA that is only passed down the maternal lineage (Hutchison et al., 1974)

individual actually inherits came from, they may harbour the same sequence (that is, the individual is homozygous in a given base or gene) or may be different between those so-called alleles (that is, the individual is heterozygous). This together with our environment is what makes us look different, but also what may cause us to be more or less susceptible to a certain disease or treatment thereof. Since the genes were assembled in the fertilized egg already that afterwards underwent cell divisions (in a process called mitosis), all the cells in our body in theory harbour the same DNA sequence, with the exception that sex cells only contain one copy that was assembled in a cut-and-paste process (called meiosis) of the two somatic ones (Hotta, Ito, and Stern, 1966).

Somatic variation

Should we sequence each cell in an individual, we would not find that all of their consecutive bases made up of As, Ts, Cs, and Gs are indeed identical, but there is also variation within each individual. There may be a change that is common to cells that derive from the same parental cell, or a change may have been introduced in a single cell by an internal or external process. In the former case, we need to realize that the DNA copy mechanism (also called DNA replication) that ensures that when a cell splits both its daughter cells inherit two full copies of its DNA is not infinitely accurate. Rather, this process may introduce reading- (from the ancestral or template strand) and writing (the newly synthesised strand) errors. In fact, in each cell division there is an average of 100 errors that are introduced while copying the each of the three billion bases (times the two alleles) that is our genome (T A Kunkel and Bebenek, 2000; McCulloch and Thomas A Kunkel, 2008).

Such an error may manifest itself in multiple ways. The replication apparatus (DNA polymerase) may miss the insertion of a base or a couple of bases that was in the ancestral in the newly synthesised strand, which produces a small deletion, or insert a base that was not in the original strand, thereby producing an insertion. It may also insert the wrong base, which leads to a substitution (Loeb, Springgate, and Battula, 1974). While discussing these mutational processes one should keep in mind that those errors introduced do manifest itself only on one of the two strands in an allele, which will lead to a base no longer matching its opposite pair as well as if there had been no error.

Mutations, copy number alterations, and structural variation

Even when the replication process produced a perfect copy of the ancestral strands, there may still be cell-internal or external processes that affect the DNA's integrity, *e.g.* exposure to ultraviolet radiation,

or a host's virus defence system (Alexandrov et al., 2013). This could lead to single base exchanges (single nucleotide polymorphisms, SNPs), small insertions and deletions (indels), changes in the number of copies of DNA segments (copy number alterations or aberrations, CNAs–this can happen either on the genome or on extragenomic small chromosomes (D. T. W. Jones et al., 2012)), or strand breaking and rejoining at different positions (structural rearrangements–*e.g.* forming gene fusions (Nowell and PC, 1960; Mitelman, Mertens, and Johansson, 1997; Garnett et al., 2012)).

1.1.3 *Functional impact of mutations: the central dogma*

A gene that has an erroneous sequence somewhere on the DNA is, by itself, not a cause for a cell to alter its function. Instead, genes only store the information, like having a template, that is required to form an active compound from it. Genes can be transcribed in regulatory RNA, or messenger RNA (mRNA; both derived from DNA by a process called transcription) that is later used by the Ribosome to chain amino acids together to a functional protein in a process called translation. The process of transcription and translation is referred to the central dogma of molecular biology (Crick, 1958).

The path from mRNA to protein may involve additional steps, such as cutting out unneeded parts or pasting together different building blocks (splicing), or covalently attaching sugars or other chemical groups (post-translational modifications) to yield the functional protein and/or to regulate its activity (Crick, 1958). Proteins are involved in the transduction of a signal from a molecular cue binding to a receptor on a cell's surface is how a cell responds to changes in its environment. In turn, this signal is relayed through the network of kinases and phosphatases[2] until it reaches the terminal nodes that are the transcription factors, acting in conjunction with polymerase III and other co-factors to initiate changes in gene expression. Those genes are in turn transcribed into RNA, and if they are protein-coding consequently translated into those.

Take, for instance, a protein involved in relaying a signal from the cell surface, such as a signal to grow and divide, to another protein that in turn activates some other proteins which ultimately induces the expression of genes that are needed for the cell's replication machinery. These signals are usually relayed by post-translational modifications (PTMs) from one protein to another. One of the best-studied processes is phosphorylation (the addition of a phosphate group - proteins that do that are called kinases, the ones that remove phosphatases) of Serines or Threonines. These are amino acids with an OH group in their side

2 there are other chemical modifications, but those are the best studied

chain and are thus susceptible to it (Burnett and Kennedy, 1954). This protein now may be affected by a mutation so that it no longer waits for an upstream signal to transduce, but sends the downstream signal to grow and divide irrespective of what the upstream input is. One such example is the well-known $BRAF^{V600E}$ mutation, where a Valine on position 600 (the 600th amino acid) is replaced by Glutamic Acid (Zecchin et al., 2013).

Genes that are relevant for cancer development and progression have long been classified in Oncogenes and Tumour Suppressor Genes (Croce, 2008). In broad terms, this classification represents the following: does a mutation introduce a gene function that was not there before, or it loses sensitivity to an inhibitory signal that kept it in check, this gene is called an Oncogene in its active (mutated) form, or Proto-Oncogene in its wild-type form; conversely, if a mutation causes a gene to lose its function as e.g. a negative regulator of cell growth, it is called a Tumour Suppressor Gene. The first Tumour Suppressor Gene that has been found was *TP*53 in 1979, now termed "guardian of the genome," whose protein is involved in numerous functions like maintaining DNA integrity, managing DNA repair, or causing apoptosis or senescence if the former fail (Levine and Oren, 2009). More recently, mutations that actively contribute to either development or progression of a cancer have been called drivers, while the mutations introduced by e.g. a faulty DNA replication machinery (that a cancer may have caused) that bear no functional impact on the cell are called passenger mutations. Driver genes are usually either activated Oncogenes or Tumour suppressor genes, but may also have both functions (Vogelstein and Kinzler, 2004).

1.1.4 *Malignant transformation*

Mutagenic driving forces (Alexandrov et al., 2013) are not only all around but also inside us (as exemplified by the *APOBEC* virus defence). However, most cells do not start to divide uncontrollably, even if they acquire a driver mutation. For instance, it has been shown that in healthy skin there is a high number of potential driver mutation that pre-exist without cancer ever developing from them (Martincorena et al., 2015). For a cell to develop into a malignant tumour, it requires multiple "hits" that transform it into a truly malignant state. The mutations and mechanisms but which it acquires those properties are different from cancer to cancer, yet the biological processes it needs to modify are remarkable similar (Hanahan and Weinberg, 2000; Hanahan and Coussens, 2012):

- Self-sufficiency in growth signals: being able to grow and divide without external cues present

- Insensitivity to anti-growth signals: being able to grow despite usual inhibitory mechanisms
- Evading apoptosis: eliminating checks and bounds that make an abnormal cell commit suicide
- Limitless replicative potential: replenishing telomeres that grow shorter after each division
- Sustained angiogenesis: promoting blood vessel growth to get an increased supply in nutrients
- Tissue invasion and metastasis: breaking the local confinement of the tissue it first arose
- Avoiding immune destruction: not being targeted by cytotoxic T-lymphocytes
- Deregulating cellular energetics: still producing energy with a lack of oxygen

1.1.5 *Tumour heterogeneity*

Cells in a tumour are not homogeneous, but the latter should rather be seen as an evolutionary process of diverse clones that compete for a growth advantage subjected to positive and negative selection. This has implications for cancer growth as well as therapy, and hence assessing this heterogeneity has become an important research topic (Heppner and Miller, 1983; Alizadeh et al., 2015).

However, there are a couple of challenges when trying to assess the whole genetic diversity of a sample. It is important to take it not only in one place, but either homogenise the tissue or selectively take small samples in different subsections. Another challenge is that for detecting very minor frequencies, one needs to read the genome many times over and be sure that the resulting variants are not only due to errors in this process, or any other by which the sample is treated after acquisition.

1.2 THERAPEUTIC INTERVENTIONS

1.2.1 *The therapeutic window*

Within a therapeutic intervention, our goal is to selectively treat (or kill) the diseased cells while not impacting the normal function of other cells in the body. If at first we assume that a treatment can be delivered to all cells of the body in the same amount, the question is whether the impact it has is impacting the cells we want to act on more than

all the other cells we want to impact as little as possible. If we take the simple example of killing cells, we want a drug that kills all the bad cells and leaves the good cells alone. Hence, an effective drug will work on the bad cells at a lower concentration than on the good ones. If the concentration is too high, it will undoubtedly affect other cells as well. This range of concentration, where a given drug already acts on the target cells but does not confer toxicity to other cells is called the therapeutic window. The broader it is, the easier it is to work with this drug.

Elaborating on our simplification, it is of course not the case that an administered compound will be available to all cells at the same concentration. Instead (Ruiz-Garcia et al., 2008), it first needs to reach the bloodstream (which includes uptake and resorption, then modification in the liver and pancreas if taken orally), then be distributed by the flow of blood into all the capillaries (with the exception of the brain that has an additional barrier), until finally cells are able to absorb it. As a rule of thumb whether a drug can be orally absorbed or not, a common measure is "Lipinsky's Rule of Fives" (Lipinski et al., 2012), that states that an orally effective drug has no more than one violation of (1) no more than five hydrogen bond donors, (2) no more than ten hydrogen bind acceptors, (3) a molecular mass less than 500 daltons, (4) an octanol-water distribution coefficient ($logP$) of less than five.[3]

1.2.2 *Cytotoxic drugs*

The easiest way to kill cancer cells is to expose them to a compound that is generally toxic to cells. While this is expected to have a negative impact on all cells, there are certain properties of cancer cells that make them more susceptible. One such property is that they grow more actively than most other cells, and so a therapeutic window exists for compounds that interfere with cell growth, as is the case for most chemotherapy. This could be DNA synthesis (Cisplatin or Carboplatin), microtubule disassembly (Paclitaxel and more generally Taxol-based compounds), or others (Skeel and Khleif, 2011).

1.2.3 *Targeted therapies*

A recurrent theme of cancer development and progression is the aberrant activation of specific molecular cues in cell signalling. An example of this is the well-known $BRAF^{V600E}$ mutation, already mentioned in section 1.1.3. But there are many more instances where a mutation in a gene yields a gene product that is abnormally active or inactive. This

3 Note that these are only four and not five rules. The "five" in the rule's name stems from the components being multiples of five, not that there are five rules

often includes members of the MAP kinase pathway (like EGFR, RAS, RAF, MEK, or ERK), but frequently also other proteins and pathways that confer a selective advantage in a given context. An important aspect in that regard is the concept of oncogene addiction: Once a cell, or a population of cells, suffers a molecular lesion that causes aberrant signalling, the cell (or cells) become dependent on exactly that signal. Turn it off, and those cells are likely to die (Weinstein, 2002).

This has an important implication for cancer therapy: if we are able to find (or design) a compound that specifically turns off one of those abnormally active proteins, we can kill cells that have it active. In example of the $BRAF^{V600E}$ mutation, computational chemists indeed designed and developed a small-molecule inhibitor called Plexicon (more specifically, PLX4720 or nowadays called Vemurafenib) that proved to bind the mutated version of the BRAF protein with a much higher affinity than the one found in wild-type cells. This, in some ways, could be considered the perfect drug for cancer therapy, because it has favourable uptake in the body and afterwards the therapeutic window is much larger compared to other compounds due to this difference in binding affinity (Chapman et al., 2011).

There are, however, many other kinase inhibitors as well as antibodies that specifically target a protein to abrogate its activity.

1.2.4 *Development of resistance*

Unfortunately, targeted therapies often can not kill all cells before they acquire a resistance mechanism to the treatment, or a sub-population of cells were resistant to begin with. These cells then outgrow their competitors. This is one of the reasons why targeting mutations in driver genes is a good start, but it needs to be augmented with knowledge of the dynamic changes mutations induce in the cellular signalling network and the evolutionary paths for a cell to respond.

Let us revisit one of the most well-known examples in targeted therapies, the inhibition of $BRAF^{V600E}$ mutants. This example of targeted therapy works with the intended effect in such that it kills cancer cells to an extent that makes multiple, from the outside clearly visible, tumours completely disappear for months (Chapman et al., 2011), which could be hailed as a success for identifying a target and rationally designing an inhibitor that abrogates oncogenic signalling. However, there was one drawback: after a couple more weeks past the initial success the tumour cells were able to overcome the effects of the inhibitor, leading to the recurrence of tumours that the patient later died from.

This, together with multiple other examples, proved to show that targeted therapies work for a while, but they are, in a lot of cases, unable to permanently suppress tumour growth. But how does this

work? In order for a tumour to regrow, there need to be some cancerous cells left alive after treatment, as this rapid regrowth of mass can not be explained by an independent inception. Hence, some cells must have survived the process of therapy (Heppner and Miller, 1983; Reya et al., 2001).

Many distinct strategies have been suggested to more effectively kill cancer cells, most notably the combination of a MEK and BRAF inhibitor in melanoma (Long et al., 2014). In fact, the very nature of killing off the majority of cells and pushing the rest through an evolutionary bottleneck has been suggested as a basis for designing therapies taking into account those temporal patterns in adaptation (Hata et al., 2016).

1.3 DISEASE MODELS

1.3.1 *Rationale*

It is easy to argue that in order to find better cancer treatments, we first need to understand better the molecular mechanisms that drive it. This is especially true because the mechanisms involved from the inception of cancer up the its progression and spreading are in fact molecular mechanisms that have spun out of the normal biological control that cells of an organism exert on each other. In order to improve our understanding, it is required to collect data about the different types, stages, and treatments of the disease.

This poses a problem: it is not possible to perform all of the assays required on the actual patients. There is just no justification for a patient to undergo surgery if we want to know if protein A interacts with protein B. Also, we can not try new treatments without a strong indication that this might be the best known possibility of curing a patient, as this would be highly unethical. It follows that there is a requirement for some biological system that mirrors the disease while, at the same time, is easy to handle in the laboratory. In choosing such a system, the points mentioned involve a trade-off: one that mirrors the disease perfectly and is easy to handle does not exist, but there are multiple systems (outlined in the sections below) that are closer to one or the other. Which one to use must be decided in each experiment individually - considering its goals, effort, conclusiveness and applicability to the question studied.

If these disease models provide a strong indication that a certain treatment approach is truly beneficial to a certain patient or patient group, this needs to be thoroughly tested to make sure that those indications previously shown using a model system - that are known to mirror a lot of aspects of the actual disease, but never all - actually hold in humans as well (Fields and Johnston, 2005).

1.3.2 *Cell lines*

One of the easiest model systems that provide a reasonably accurate mirror of a lot of the biology involved in the disease are cell lines, which is basically taking a number of cells from a tumour, and growing them in a petri dish for multiple generations (passages). An advantage of this method is that the biological material that assays can be performed on is virtually unlimited, as well as easy to produce and maintain. This is because cells grow in the dishes as long as they are supplied with nutrients and potentially growth factors or cytokines, while still reflecting a lot of the properties that cells in a tumour would. However, by passaging cells consecutively in petri dishes for many generations, the evolutionary pressure that they are selected by is essentially the growth rate on the dish, which will not reflect the forces one would find in a real tumour. In fact, the fastest growing clone will outcompete all others in a couple of generations - which gives a relatively uniform molecular phenotype across all cells, but in turn also loses the heterogeneity found in a primary sample.

Today, there is a multitude of cell lines available from commercial suppliers. The maybe best-known example is that of the HeLa cell line (Gey, Coffman, and Kubicek, 1952) that was derived from a patient of cervical cancer in 1951, named Henrietta Lacks. This cell line, however, considering how long it has been kept in culture as well as its high mutation rate, has produced a genotype that no longer resembles primary cancer samples in many ways. One property is that a normal human cell has two copies of the genome (the two alleles), whereas HeLa cells have four copies, and its genome sequence revealed more extensive aberrations (Landry et al., 2013). In most cases however, experimentally used cell lines are more closely tied their origin: they have, albeit with aberrations, a genome that resembles primary tumours and they hence in many ways still resemble their tissue of origin in molecular terms (Iorio, Knijnenburg, et al., 2016).

However, there are biological dynamics involved in a real tumour that can not be recapitulated by cell lines, like their interactions with other cells (especially immune cells) or with their surroundings (the extracellular matrix). Also, their number is limited. With the approximately 1,000 cell lines that the Genomics of Drug Sensitivity in Cancer (GDSC, cf. section 1.5.4) screened for compound sensitivity (Iorio, Knijnenburg, et al., 2016), it is time to think about how many more cell lines would be needed to gain enough statistical power for the high number of rare mutations and whether the required amount of diversity can theoretically be generated with cell lines (Francies and Garnett, 2015). One way to resolve this issue could be organoid cultures: these are small assemblies of primary cells taken out of a tumour and grown in

matrigel, where the spatial separation of mini-cultures retains much of the tumour heterogeneity (Sachs and Clevers, 2014). Because of these properties, they could advance *in silico* cancer screening in a way that is not possible with other platforms (Francies and Garnett, 2015).

1.3.3 *Animal models*

For some kinds of experiments, it is necessary to model the effects on a whole organism. For this purpose animals have long been used, from a simple multicellular worm (*Caenorhabdilis elegans*, mostly for development and its simple central nervous system) to frogs (*Xenopus*, for induced pluripotency), and mammals (mice, rats, and chimps; often for diseases). Just as the complexity of these organisms increases, they also mirror more closely aspects of human physiology. However, experiments performed in higher organisms also require more time, they are more difficult (and expensive) to set up, and there are ethical concerns on keeping, handling, and killing animals for the purpose of ultimately saving human lives.

In terms of cancer, looking at the number of articles published about a specific molecular mechanism or the detailed characterisation of a compound and its usability as a drug, mice seem to be considered a reasonable trade-off between providing a close enough match of human biology and being not too difficult to keep (Fields and Johnston, 2005). To gain insights, a mouse's genome can be engineered to include for instance a mutated version of a human gene that is known to induce the formation of tumours (Talmadge et al., 2007), or can be inoculated with cancerous cells derived from either a cell line or a patient (Fogh, 2014). Mouse models have taught us how cancerous cells influence and modify the microenvironment around them and how the cells can sustain themselves and grow, attract the formation of blood vessels, and ultimately metastasise (Talmadge et al., 2007).

1.4 MOLECULAR DATA TYPES AND ASSAYS

1.4.1 *Role of molecular data*

One of the issues of both understanding as well as preventing or treating cancer development and progression is that it is a process with such complexity that a simple observation of patients with the disease will not suffice to come up with effective models of disease inception and progression, or treatments for that matter. Fortunately, we have got a battery of tests available, quantifying different molecular aspects of a biological sample. Examples of these data include the sequence, structure, and modification of DNA, but also the expression of RNA

or proteins, including modifications. These different layers, able to quantify different projections of a cell's state, have provided us with an unprecedented opportunity to truly understand many molecular mechanisms that govern the processes active in both a normal and a diseased cell, and the differences between them.

1.4.2 *DNA*

When Frederick Sanger discovered (Sanger, Nicklen, and Coulson, 1977) that one could read DNA by supplying a minor fraction of di-deoxynucleotides in addition to labelled deoxynucleotides that are normally incorporated into DNA during replication, in turn halting the replication chain and revealing the sequence, it was soon clear that this technology would revolutionise biology and medicine. Further incremental improvements and then the switch to next-generation sequencing has enabled the scientific community to read DNA on an unprecedented scale (Schuster, 2008), with the implication that the genome of individual (cancer) patients could soon be used to decide upon optimal treatment of each individual (McDermott, 2015).

1.4.3 *RNA*

All healthy cells contain roughly the same genome, yet the functions they perform in the body is vastly different. This is why in contrast to DNA, the RNA that is transcribed represents more the state a cell is currently in and functions that need to be performed at a given time and in a given context. It can hence be argued that the sum of RNAs (the transcriptome), or more specifically mRNAs, provides an overview of that state. Measuring those RNAs will allow us to get a picture of the processes going on in a cell at a given time. The two ways by which this is usually done are either microarrays or RNA sequencing, as outlined below.

Microarrays

The classic way to measure RNA expression is using microarrays (Brown and Botstein, 1999; Debouck and Goodfellow, 1999). These are, in the simplest case, a carrier matrix (glass can be used but the more recent chips use a polymer matrix) that has fixed single-stranded oligonucleotides spotted on them. From our sample, we would then extract and clean the RNA while simultaneously degrading any DNA that it contains, and label the RNA either with one dye (Gohlmann and Talloen, 2009; Du, Kibbe, and Lin, 2008) (if we want to quantify the expression levels of transcripts in a given sample) or two different dyes

(Shalon, S. J. Smith, and Brown, 1996) (if we want to quantify the difference between two samples). Such fluorescent dyes used are e.g. Cy3 and Cy5, that emit light in the green and red wave lengths upon stimulation, respectively. Note that if the technical reproducibility of the platform is good enough, it is also possible to compare conditions by using two different arrays where we measure each sample individually and then compare the outcome. - A technique that has been favoured by companies like Affymetrix as it allowed chips at much higher density and the comparison between each sample in a control- and in a experimental condition, as well as between them.

An obvious drawback of this technology is that we need to know in advance which oligonucleotides to spot, so in turn which genes we are looking for. There needs to be different chips for different organisms, depending on the gene sequences that they carry. This selection of genes that we look at can of course introduce bias, in such that we can't look for transcripts that we don't know exist, but also we might focus on transcripts and isoforms that we think are important before carrying out an experiment. Another possible issue is their detection threshold. As the readout is fluorescence based and there will always be a base-line level of it, and it is hard to quantify the amount of RNA bound on a spot of the total number of molecules are so low that a few molecules do not significantly change the readout. Turning this argument around, RNA binding to spots also has its point of saturation, where it does not matter if there is more RNA present or not once all probes bind to their complementary sequence and thus produce a signal (Duggan et al., 1999).

RNA sequencing

An alternative approach that has become more popular recently with the falling costs of DNA sequencing is to apply this technology to RNA as well (Marioni et al., 2008; Z. Wang, Gerstein, and Snyder, 2009). In this case, we can use the RNA that we isolated from a sample, reverse-transcribe it into DNA, and sequence the DNA like we would normally. This has got the advantage that it can be used even if there is no genome sequence or chip available and as a consequence we are also not biasing our selection of genes by previously knowing what to look at (Trapnell, Williams, et al., 2010), with the exception of sections in the RNA that can be more less easily transcribed into DNA (like GC-rich regions that also pose challenges for sequencing). Another advantage is that the dynamic range of this technology (*i.e.*, the differences in transcript numbers it can detect) is much higher than it is for microarrays, both on the lower as well as on the upper bound: we can detect a single read, but also having a lot of reads won't saturate our signal the same way that spots on the microarray do (although if a large proportion of the

reads is from a highly abundant molecule, this will decrease sensitivity for detecting other RNAs).

The way RNA-seq quantification (Mortazavi et al., 2008; B. Li and Dewey, 2011) is usually done is to align the reads to a known genome, allowing gaps in the alignment for splicing and differential use of exons (Trapnell, Roberts, et al., 2012). The transcription level of a gene, transcript or exon are then quantified by how many reads align to the corresponding section in the genome. As this is computationally very expensive, there have been efforts recently to get around performing the alignment step by k-mer counting and so-called pseudo-alignments. Methods for the latter include Sailfish (Patro, Mount, and Kingsford, 2014), Salmon[4], or Kallisto (Bray et al., 2016).

The data produced by RNA-sequencing is different to the one from microarrays, as it is based on read numbers that thus counts (so discrete data) instead of the continuous fluorescence signal that microarrays provide. These counts are negatively binomially distributed (Simon Anders and Wolfgang Huber, 2010), and require specialised software packages to call e.g. differential expression like *edgeR* (Robinson, McCarthy, and Gordon K Smyth, 2010), *DEseq* (Simon Anders and Wolfgang Huber, 2010) and *DEseq2* (M. I. Love, Wolfgang Huber, and Simon Anders, 2014; M. Love, Anders, and Huber, 2014), or transformation before they can be used in standard linear modelling techniques, for instance provided by `voom` in the *limma* package (Law et al., 2014).

1.4.4 *Proteins and phosphorylation*

It can easily be argued that in order to get an appropriate picture of what goes on in a cell it would be better to look at the proteins that perform most functions, as opposed to the mRNA levels of genes. The latter will in many cases be translated into proteins that then exert their activity, yet mRNA measurements are one more step removed from the functional process than proteins, and two steps from measuring post-translational modifications (PTMs) that modify activity already gives us a lot of functional information - as long as we know how to interpret it.

Taking these facts together, we might want to look at proteins instead of gene expression. Yet, gene expression is a lot cheaper and easier to measure and provides more coverage than the proteomic methods currently can. In addition, the publicly available gene expression data far exceeds the one of proteomic data. There are different experimental methods for quantifying proteins and their state, the most well-known are listed below. This thesis, however, is focussed on gene expression for the reasons outlined.

4 https://github.com/COMBINE-lab/salmon

Reverse-Phase Protein Arrays (RPPA)

In this method (Tibes et al., 2006), there is an array of spotted antibodies, similar to the nucleic acids on a microarray. A labelled cell lysate (or other sample) is incubated with the antibodies, and we can quantify the amount of the proteins (or phospho-proteins) corresponding to each antibody afterwards. In contrast to microarrays, however, this is relatively low throughput as antibodies can not be as readily synthesized as nucleic acids.

Mass spectrometry

Mass spectrometry (Domon and Aebersold, 2006; Bensimon, Heck, and Aebersold, 2012) is based on fragmenting proteins into peptides that are then ionized and sprayed into an electrical field in a vacuum tube, where their mass-to-charge ratio causes them to behave in a certain way. For instance, in Time of Flight (TOF) instruments, the electrical field is used to accelerate the peptides; the force that accelerates them is then proportional to their charge, and their inertia to their mass– hence the time it takes to reaches the detector is proportional to the ratio of the two.

For more complex samples, there are too many peptides to detect at any given time, so the Mass Spectrometry is often coupled with another technology that first separates proteins contained in a sample using some other property (e.g. their hydrophilicity/hydrophobicity) in columns of gas (gas chromatography, GC) or liquid (high performance liquid chromatography, HPLC).

1.5 DATA SETS

1.5.1 *Public gene expression repositories*

There are two major repositories of gene expression experiments using microarrays, the Gene Expression Omnibus (GEO) (Barrett, Troup, Wilhite, Ledoux, Rudnev, Evangelista, I. F. Kim, Soboleva, Tomashevsky, and Edgar, 2007; Barrett, Troup, Wilhite, Ledoux, Rudnev, Evangelista, I. F. Kim, Soboleva, Tomashevsky, Marshall, et al., 2009) at the National Institutes of Biotechnology Information (NCBI) and ArrayExpress (Parkinson et al., 2007; Helen Parkinson et al., 2009) at the European Bioinformatics institute. The reporting standards defined in the Minimal Information about a Microarray Experiment (MIAME) (Brazma et al., 2001), played a pivotal role to ensure that experiments where one could not only measure one gene but the entire transcriptome remained interpretable. Both of these are synchronised, which means that submissions one will be imported and made available on

the other as well. More recently, ArrayExpress also started collecting RNA-seq and ChIP-seq experiments.

1.5.2 *Gene sets and pathways*

With about 22,000 human genes whose transcription can be measured simultaneously using the methods described in section 1.4.3, it is important to put genes into functional groups for several reasons: (1) groups of higher level processes are more interpretable, (2) gene level measurements in microarrays (and to a lesser extent RNA-sequencing) are inherently noisy and the more measurements we combine the clearer the signal gets, and (3) it may be possible to solve statistical problems where the number of observations per sample is greater than the number of samples.

There are multiple databases available that link gene sets into function groups. Some of the most well-known are listed below, but there are others.

Gene Ontology

Gene Ontology (GO) (Ashburner et al., 2000; Gene Ontology Consortium, 2004) is "a major bioinformatics initiative to develop a computational representation of our evolving knowledge of how genes encode biological functions at the molecular, cellular and tissue system levels". It has encoded over 40,000 biological concepts based on experiments reported in over 100,000 publications according to its web portal at http://geneontology.org/.

KEGG

KEGG (Kanehisa and Goto, 2000) was one of the first pathway databases. To provide an idea of how widely it was (and still is) used, Google Scholar (query 25th February 2016) reported 6863 publications that cited the original article. However, in 2011 the platform went commercial, allowing access only via a subscription-based portal.[5] While the authors removed public access also to earlier versions of the database, its pre-commercial license (OICR by Pathway Solutions, Inc.) allowed it to still be used. It will, however, not receive any more updates.

Reactome

Reactome (Croft et al., 2011) is a "a free, open-source, curated and peer reviewed pathway database" hosted and curated (mostly) by the

5 http://www.pathway.jp/

European Bioinformatics Institute (EMBL-EBI). It has a web portal at http://www.reactome.org/, and the current version is v55 released in December, 2015. It has an open license (Creative Commons Attribution 4.0[6]) for its own data and the pre-commercial license for imported KEGG data.

1.5.3 *Primary cancer cohorts*

A lot of molecular information about cancer cohorts has been collected, analysed, and released. The best-known are the TCGA and ICGC.

The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas is a United States-based consortium that aims to profile primary tumours belonging to many different cohorts on the DNA, RNA, protein, and epigenetic level. The initial release (The Cancer Genome Atlas Research Network et al., 2013) contained 12 tumour types, but it has grown to 35 with their June 2016 release. There are many secondary portals that allow access to the TCGA data, such as the BROAD Firehose tool[7] or cBioPortal (Gao et al., 2013), the latter of which also includes data from other projects.

International Cancer Genome Consortium (ICGC)

The International Cancer Genome consortium is an umbrella effort with the same goals, yet it incorporates many more studies from different countries and thus represents a superset of the TCGA data (International Cancer Genome Consortium et al., 2010).

1.5.4 *Cancer cell line drug sensitivity resources*

There are many studies measuring the drug sensitivity of different cell lines. The two biggest are the GDSC and CCLE.

Genomics of Drug Sensitivity in Cancer (GDSC)

The first release of the cell ling drug screening from the Cancer Genome Project of the Wellcome Trust Sanger Institute contained 639 cell lines treated with 130 cancer drugs that were in either clinical use or pre-clinical development, where the cell lines have also been profiled by capillary sequencing of 77 known oncogenes (Garnett et al., 2012). The subsequent release increased this count to 1,001 cell lines and 265 drugs,

6 https://creativecommons.org/licenses/by/4.0/
7 https://gdac.broadinstitute.org/

with a full molecular characterisation of cell lines comprised of full exome sequencing, SNP6 arrays for copy number, high-quality microarray gene expression, and DNA promoter methylation (Iorio, Knijnenburg, et al., 2016).

Cancer Cell Line Encyclopedia (CCLE)

The Cancer Cell Line Encyclopedia (Barretina et al., 2012) characterised the mutations of 947 human cancer cell lines along with SNP 6.0 copy number and Affymetrix U133 Plus 2.0 array gene expression, approximately 500 of which they treated with 24 anti-cancer compounds.

In contrast to the GDSC, they did not chose concentration ranges of the screened drugs in order to detect the few cell lines that are sensitive to a given drug, but rather a generally applicable range. The resulting differences compared the the GDSC caused a number of inconsistencies (Haibe-Kains et al., 2012) but they are largely resolved (Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium, 2015).

1.5.5 *Drug-induced transcriptional changes*

Original Connectivity Map using microarrays

The first large-scale project providing signatures of drug-perturbed gene expression changes in the MCF-7 cell line was the Connectivity Map (Lamb et al., 2006). It comprised a total of 1,309 compounds, yet most of them were not anti-cancer compounds. It has introduced and enabled the paradigm of signature matching (section 1.6.3), a method of using gene expression changes upon drug treatment to match either drugs with drugs for finding similarities or drugs with diseases for a potential treatment indication (Iorio, Rittman, et al., 2013).

The L1000 platform

The new version of the Connectivity Map is based on Luminex beads (Peck et al., 2006) that are able to measure about 500 transcripts. For the L1000 platform[8], the BROAD institute managed to put twice the amount of genes on one bead by scanning it in two different dilution ranges and deconvoluting them computationally afterwards. After that, they scaled the experimental readout by 80 control genes that are supposed to be constant across experiments, optionally infer the whole transcriptome from their 978 "landmark" genes (projections shown in figure 1), and compute z-scores (number of standard deviations of a

8 More information is available at: http://www.lincscloud.org/l1000/

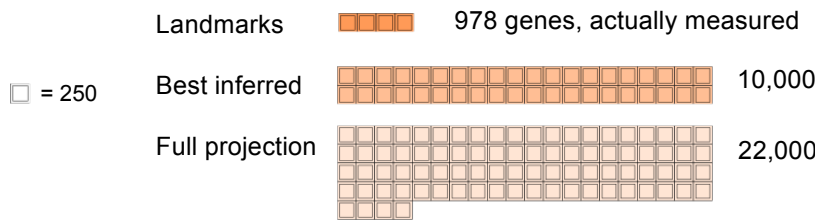

Figure 1: Overview of number of experiments available in LINCS using cancer drugs.

perturbed condition over the mean of the control) for each perturbation[9].

In 2015, the BROAD institute released the raw data and z-scores between each control- and drug-perturbed experiment for 978 oligonucleotide probes and a total of 1.4 million conditions in a 111 gigabyte *.gctx* (HDF5 and metadata) file where they projected their actual measurement to the full gene space using publicly available microarray data. At this time, no one knew about the quality of the data (the authors claimed it to be equal with microarrays; my own tries and conversations with other people using the data have indicated that it is below that). The main publication is still not out today (August 2016), and they refused to share *e.g.* the linear transformation matrix they used to obtain the projected gene space.

1.6 COMPUTATIONAL METHODS

Over the last decades and years, biology moved from an observational, qualitative to a very much quantitative science. This can largely be attributed to high throughput assays, led by the advance of DNA sequencing and followed by its derivatives as well as other approaches, like high-content phenotypic screening - each allowing for measuring not one data point, but hundreds and sometimes thousands at a time. The magic component in this process is automation: the bulk of the data generated by scientific laboratories are not longer carried out by individual scientists pipetting together reagents, but by machines controlled by computers and designed by both scientists and industry alike. This shift of paradigm has not only produced much more data, but arguably led to an increase in the quality of such data as well, by removing the human element (mood, ability to concentrate, etc.) from large parts of the outcome of an experiment.

This transformation has not only come with its benefits, but also with its challenges. Setting aside the view of some prominent scientists that each experiment needs to have a specific question or hypothesis in mind and that generating data first and using the data itself as a hypo-

9 http://www.lincsproject.org/data/data-releases/

thesis generator is an inferior approach (as best demonstrated by Sidney Brenner's quote "low input, high throughput, not output" (Friedberg, 2008)), a biomedical scientist's skill requirements have moved from knowing, for instance, as many details about a given gene as possible to understanding how to set up reproducible assays, generate reliable data, but especially treating the resulting data in a way that yields biological insights.

There are two crucial parts to this: one is algorithms, that transform the raw data generated by a given experiment into biologically interpretable indications, and the second is statistics, to make sure the effect observed is due to the data and not due to random chance while analysing the results. I will outline some very successful ideas that in turn led to algorithms and usable software packages for investigating molecular data, both in terms of cancer as well as in general, in the sections of this and the next section.

However, let us first reconsider the goal we are trying to achieve in terms of this thesis as well as a lot of the related work: we ultimately want to improve how cancer is diagnosed and treated for patients, and those algorithms help us to define markers of their pathogenesis or how they could reap benefits from being offered a particular treatment (Rubio-Perez et al., 2015). We thus want to transform the numbers reflecting the molecular data we have in a meaningful way so they connect to this endeavour.

1.6.1 *Patterns of mutations*

When looking at sets of mutations instead of individual ones distinct patterns emerge. One such pattern is that there are modules that are either co-occurring or mutually exclusive (Babur et al., 2015). In the event of co-occurring mutations, this might mean that a given mutation is not sufficient to obtain a certain trait and thus a second mutation is necessary to confer it. In the case of mutual exclusivity, a second mutation is not conferring a growth advantage after the first one arose. This may be due to evolutionary parsimony or a fitness defect. In the first case, a second mutation that inactivates, for example, an already inactivated tumour suppressor pathway is unlikely to happen on the population-level because there is no selective pressure, as the required trait has already been acquired by the cell. In the second case, the growth advantage conferred by one mutation might be cancelled out or counteracted by the presence of a second one, thereby mediating a selection growth disadvantage (fitness defect) of the cells carrying both mutations as opposed to either one of them.

Much focus has devoted to identifying driver mutations in different cancer types that, combined with oncogene addiction[10], forms the basis of many targeted therapies: if you can inhibit signalling stemming from a mutation that drives a cancer's development and progression, the affected cells' growth will be severely abrogated.

The concept that a second mutation kills a cell with a given mutation has been termed synthetic lethality. It has been used to study genetic interactions in model organisms for a long time (Nijman, 2011). The loss of both genes in a synthetic lethal pair is, as the name suggests, lethal to the organism, but the loss of each individual gene is not. More recently, these interactions have been used to identify probes interfering with RNA that are purified from a starting population (Cheung et al., 2011) and efficient computational algorithms have been developed to find those pairs in primary cancer data sets like the TCGA (Ciriello et al., 2012; Gobbi et al., 2014). These interactions can then be transformed in a network (Jerby-Arnon et al., 2014), which can then be used to predict sensitivity of cancer cell lines to a certain drug treatment, as well as clinical outcome given the level of co-expression of its pairs. An example of a synthetic lethal pair that can be therapeutically exploited are loss of function mutations in *BRCA*1/2 and treatment with PARP inhibitors (Farmer et al., 2005).

Including those sorts of analyses is an interest and future goal in my research, but has not contributed significant results to this thesis.

1.6.2 *Gene expression clustering*

Cancer is known to be a heterogeneous disease, with individual tumours forming different subtypes even if it arose from the same tissue. One of the arguably most effective ways of elucidating the state of individual tumours and cell populations therein is quantifying its transcriptome using methods such as microarrays or RNA sequencing.

We can use the global pattern of gene expression to identify different subtypes of the disease in a given tissue (or even between tissues). A challenge with this is that the transcriptome could be comprised of up to 22,000 genes without even considering most regulatory or ribosomal RNAs. This space needs to be reduced to something more manageable for inspection, which may reveal subgroups that have prognostic or therapeutic relevance. Approaches that allow for visual inspection of a high-dimensional data set are called dimensionality reduction techniques, while methods assigning the different samples to different subtypes are called clustering algorithms (Xu and Wunsch, 2005).

10 the tendency of a transformed cell to become dependent on sustained impact of the lesion it first obtained, see section 1.2.3

Visualising patterns in high-dimensional data

Principal Component Analysis (PCA) (Wold, Esbensen, and Geladi, 1987) is one of the simplest linear transformations that rotates samples in N-dimensional space (where N is the number of observations per sample) in a way that its projection to M-dimensional space (the target dimensions - for interpretability usually two or three) maximises the variance contained in M. The rotation is calculated using a matrix factorisation that yields a unique solution. Its parameters provide the position of a sample in M space where the axes are principal components, and the latter's projection back into N called loadings (which represent how much of each original axis is contained within the new axes). It is important to note the meaning and interpretability of such a decomposition is strongly dependent on the amount of variance captured in the reduced space. PCA is closely related to Singular Value Decomposition (SVD), which has applications in dimensionality reduction as well (Wall, Rechtsteiner, and Rocha, 2003).

T-distributed Stochastic Neighbour Embedding (t-SNE) (Van der Maaten and Hinton, 2008), instead of relying on the most variable global structures, visualises local structures in a given data set: starting from a point in space (a sample), additional samples are distributed in its vicinity depending on their distance to the original sample, as well as the other points considered (that are a subset of the total number of points). This method provides much better resolution than PCA, but one can only trust the points neighbouring each other up to a number specified by the perplexity parameter. It has later been extended by the original author to use the Barnes-Hut method (Maaten, 2013) that is usually used in large N-body simulation in astrophysics (and chapter 4).

Clustering

One of the simplest clustering algorithms is K-means (Hartigan and Wong, 1979) that requires the number of clusters to be known *a priori*. It starts with assigning N cluster centres randomly in a given data set, and then assigning all samples that have a smaller distance to a given centre than the others to that centre. These assignments are iterative, which means that once the cluster centres have been determined and the samples assigned, the centres are updated to correspond to the centre of the samples each cluster is associated with. In turn, which samples are associated with which clusters is also updated after each time clusters get new samples assigned, until the process converges and further updating steps do not change cluster assignments anymore.

Non-Negative Matrix Factorisation (NMF) (D. D. Lee and Seung, 2001) is a matrix factorisation method that decomposes the matrix

V (usually with observations in rows and samples in columns) into the matrices W (with samples in columns and weights for each cluster in rows) multiplied with the matrix H (cluster assignment vectors in columns and samples in rows). It does not determine the optimal number of clusters by itself, but can be run with different numbers of clusters that is later evaluated using the cophenetic coefficient (a measure of goodness of fit of the samples to the cluster centres).

There are also other methods (Spectral Clustering (Ng, Jordan, Weiss, et al., 2002), Latent Dirichlet Allocation (Blei, Ng, and Jordan, 2003)– a modern, Bayesian method for clustering that is an active research topic. I have worked with NMF clustering to investigate cancer cell line drug sensitivity, but the results have not made it into this thesis.

1.6.3 *Gene expression signatures*

Signature definition

Maybe the simplest way to investigate a biological system using gene expression data is to not make any assumptions concerning translation, protein activity, signalling, or pathways. Instead, one could use the transcriptional state of a cell as a black box entirely and just treat it as a readout of its phenotype. This has the advantage that many of the assumptions that we normally make that are not quite correct (mRNA expression corresponds to protein level or activity, exactly those genes in a set have that function and there is no overlap or partial membership, etc.) no longer have an impact on the result we obtain - we can just use the induced gene expression of a disease, a drug treatment, or really anything else that produces gene expression changes and use it as a signature for the phenotype. These gene expression signatures can then be used for different purposes (Iorio, Rittman, et al., 2013), some of which are listed below.

Hence, we can define a gene expression signature S as a function that takes the gene expression levels E as input and transforms then into an inferred phenotype $\hat{P}$:

$$S(E) = \hat{P}$$

The signature S itself is most commonly derived from differences in gene expression E of a certain perturbation compared to a control or different sample conditions. It represents a model of transformation (*e.g.* a linear model of coefficients or GSEA on a gene list) that can then be applied to gene expression levels in a different sample E to infer the phenotype $\hat{P}$ (which is a different experiment than was used to infer E). It is important to note that the inferred phenotype $\hat{P}$ may not be the same as the actual phenotype P but is only an approximation.

Such a phenotype could be drug treatment (section 1.6.3) or pathway activity for a single condition (Bild, Yao, et al., 2005; Gatza, Lucas, et al., 2010) or using a consensus model (Parikh et al., 2010). However, both of these published approaches have drawbacks that I will discuss later.

Signature Matching

When investigating the effect small bioactive molecules (*i.e.*, drugs) have on a system, we might not want to look at the mechanism by which this happens at all, but instead rely on the changes of gene expression upon treatment (Lamb, 2007). Then we could use the downstream genes that are changing as a signature of treatment with that drug. This has got two applications that have indeed been used, which are: (1) matching a drug's signature with another drug's signature, and if they are a close match but have e.g. different indications, suggest that each drug may be used for the other indication (where directionality may be limited by additional factors), or (2) matching a drug's signature with the inverse gene expression changes that a disease exhibits over normal controls, we can suggest that this drug may be used to counteract the effects of the disease (Iorio, Tagliaferri, and Bernardo, 2009; Pacini et al., 2013), *i.e.* be a potential treatment if the disease is indeed caused by the gene expression changes that the drug reverses. However, those matches should rather be seen as hypothesis generators than definite indications for repurposing and treatment. Of course, this kind of approach relies on the availability of signatures of the drugs we look at, or of a given drug and disease, respectively.

1.6.4 *Networks*

An important consequence of the central dogma is that gene expression patterns are not randomly but hierarchically organised. Starting with perturbations on a cell's surface or its interior, a signal is propagated from its origin up to proteins that bind to DNA and change their expression as a response to the stimulus. The terminal nodes of this signal transduction are called transcription factors, that upon binding on the DNA mediate and direct (either in a promoting or in an inhibiting fashion) binding of DNA-dependent RNA polymerase (PolII) that forms a complex with available factors in order to start transcription of a factor's target genes (Watson, 1987).

The genes transcribed upon activation of transcription factors may in turn be transcription factors themselves that cause increased or decreased transcription of other genes. These interactions between different transcription factors are usually referred to as a transcription factor network or gene regulatory network (Bansal et al., 2007). There

are multiple ways these can be investigated: by characterising which transcription factors bind to which genes. This can be done either experimentally (Lachmann et al., 2010; L. Chen, Wu, and Ji, 2011; Auerbach, B. Chen, and Butte, 2013) or computationally by looking at which sequence of nucleotides a given factor is likely to bind (T L Bailey and Elkan, 1995; Timothy L Bailey et al., 2009; Matys et al., 2003; Portales-Casamar et al., 2009; Mathelier et al., 2013). An alternative approach is to find genes that change in a coordinated fashion and hypothesise that those might be regulated by the same set of factors, or by inferring which combination of which factors is required for a certain gene to be transcribed (Langfelder and Horvath, 2008; Margolin et al., 2006).

Pathways and signalling

Cell signalling is deregulated in many diseases, including cancer that I focus on here because of the sheer wealth of available data. But how to best quantify the signalling activity? The closest proxy we have would be to quantify post-translational modifications that are known to confer activity. In the simplest case this could be a phosphate group attached to a certain position that is known to make a kinase active, that is that it in turn phosphorylates its downstream targets. But phosphorylation data, and to a lesser extent protein data in general is much harder to come by than sequencing data. Mass spectrometry and Reverse Phase Protein Arrays just do not produce the same clarity that a DNA sequence or RNA level provides, plus is much harder to generate because the technology that could do it with the same throughput as sequencing technologies just does not exist.

We can thus argue that computational methods are required to make statements about cell signalling, starting from DNA and RNA instead of protein and PTM data. The simplest approach to this is looking how much mRNA of signalling molecules are expressed that comprise a pathway, and designate a high expression a high activity and vice versa (using algorithms like Gene Set Enrichment Analysis (Subramanian et al., 2005), more in chapter 2). This, however, is at odds with the way cell signalling works and is regulated. Inferring the protein levels from mRNA levels may indeed be viable (Gry et al., 2009; Maier, Güell, and Serrano, 2009), but two steps removed from the PTMs. Methods have been developed to address the issue of taking the expression level of a gene set as activity proxy by considering the structure and signs of the different pathway molecules, e.g. Signaling Pathway Impact Analysis (Tarca et al., 2008) or Pathifier (Drier, Sheffer, and Domany, 2013). They, however, do still not distinguish between expression level and activity. Another method, PARADIGM (Vaske et al., 2010) could in theory support it, depending the pathway structure supplied for

inference. Yet, the focus has never been to tell apart activity from expression. I elaborate on these issues further in chapter 2-4, with a possible way to resolve them.

Binding motifs

Some, albeit not all, transcription factors prefer to bind a specific sequence of nucleotides that they recognize on the DNA strands, called a motif. Starting from binding data such as ChIP-seq peaks, they can be found using tools such as MEME (T L Bailey and Elkan, 1995) that will report different weight matrices for nucleotides around the binding region (which correspond to how often a given nucleotide is found in a given position). These motifs in turn can be used to scan the genome for the same or a similar sequence using a sliding window approach that may indicate sites where a transcription factor could bind but did not in the original experiment. Such can for instance be the case if the chromatin is too densely packed in a region (but may not be if the cell was in a different state or from a different tissue), or the binding site is occupied by another transcription factor blocking the binding of the one we are looking at. There are databases that collect and store those motifs from experimental data and their known and predicted binding in different cell types, such as JASPAR (Portales-Casamar et al., 2009; Mathelier et al., 2013), TRANSFAC (Matys et al., 2003), or the Ensembl Regulatory Build (Zerbino et al., 2016).

Mutual Information for transcription factor networks

Another question that we might be interested in is which genes share a common regulator. In the simplest case this is a transcription factor that, upon activation, transcribes a set of genes in a coordinated fashion. Thus, we can look for genes that are expressed in a coordinated fashion across different conditions by calculating the mutual information of all gene pairs and setting a lower bound to consider. Further, we can prune the edges in the resulting network by postulating that a link between genes A and C is indirect if the mutual information of A and B, as well as B and C is higher than the one of A and C–a concept called Data Processing Inequality. This is the principle of ARACNE (Margolin et al., 2006) and Master Regulator Analysis (MRA), first used to identify the regulatory network in human B cells (Basso et al., 2005).

An extension to the method proposed by the same group is to condition the mutual information on the expression of a modulator called MINDy (K. Wang et al., 2009). The idea here is to look at the upper and lower third of the modulator expression, and calculate mutual information for both of these sets. Afterwards, instead of inferring the network, the authors look for the strongest changes in mutual informa-

tion between the two subsets: if that is the case, the modulator can be seen as a co-factor required for transcription factor regulation. If a high modulator expression correlates with an increase of mutual information it will likely be activating, or inactivating in case of a decrease.

1.7 REPRODUCIBILITY OF RESULTS

1.7.1 *Experimental and computational*

Reproducibility of scientific experiments has recently gotten into the spotlight when Amgen published a high-profile study trying to independently validate findings of 53 findings reported by different groups in the journals *Nature*, *Science* and *Cell* (Begley and Ellis, 2012; Baker, 2016). For 47 of the studies that they investigated, they could not obtain the same results that the respective authors reported. Needless to say, this raised some concerns about the current paradigm of publishing new results quickly instead of thoroughly and over-claiming the effect or significance of a study in order to have an article accepted in the top tier journals. Follow-up studies have pointed a part of the blame on use of different antibodies (Baker, 2015) that did not always show the affinity and specificity to their target as the vendors claimed that resulted in the creation of a registry for validated antibodies at (Bradbury and Plückthun, 2015).

It is easy to argue that experimental reproducibility is a hard issue to tackle, as the potential for confounding variables is huge and a lab will hardly ever have the resources to account for all of them. In general, good study design, that is proper controls and randomisation, should go a long way (especially in the case of clinical trials), but even then, there may be confounding effects unknown to the experimenters that can not properly be controlled for, as has later been shown in the case where male vs. female stewards in a mouse facility produced different reactions of the animals, potentially influencing the results obtained from a very large number of studies (S. Reardon, 2016).

Apart from experimental reproducibility, computational reproducibility should be an issue that is easier to tackle, because at least for deterministic algorithms the same input should always produce the same output using the same tools. However, this is also easier said than done because the tools will depend on other tools, maybe with different versions, and each version may change the way they treat the data to go from input to output in a slightly different way. This is especially true as computational analyses, as well as the tools employed, are getting more and more complex. However, even if the potential confounding factors are smaller than for experimental scientists, computational analysis also has its high-profile cases fraught with issues.

The most well-known example of this are maybe Anil Potti's cancer gene signatures[11] that Keith Baggerly later spent six months trying to reproduce–and failed, but in the process found considerable errors that later caused the original article to be retracted (Baggerly and Coombes, 2009).

To summarise, ensuring that a subsequent study has the possibility to arrive at the same conclusions given the same starting point, and in turn build on the results in a more confident way, has gotten a big enough issue to dedicate an introductory chapter to it. Since I did not perform any experiments to produce data myself, the focus of this chapter shall be to ensure that the results obtained by transforming raw (or in some cases already processed) data into other types of data, plots, and ultimately interpretation are reproducible in a sense that a person not involved in the projects could arrive at the same conclusion, being provided this thesis, the code used, and the technical documentation written in conjunction with it. This is especially true because the maybe best known retraction caused by computational irreproducibility handled a topic very similar to the one I investigate in this thesis: the effect of signalling pathway signatures of different cancers (chapter 3) and their significance (chapter 4).

1.7.2 *Scientific software ecosystem*

With computational analyses getting more complex, it is no longer enough to just apply a simple statistical test for a number of observations of one condition vs. another. The wealth of data that has become available needs pre-processing, normalising, statistical analysis, and interpretation of results. No one scientist can perform all of these tasks completely independently, as some of them may detailed knowledge of all the algorithms involved, up to an extent that is prohibitory. It is thus required to package repeated steps together in higher-order functionality.

This is where the scientific software ecosystem comes in. And, to a larger extent, the ecosystem of general software. For instance, I want to obtain, pre-process, and normalise microarray data to then compute differentially expressed genes in two conditions. The concepts of obtain, pre-process, and normalise are well enough defined that I should not have to worry about their exact implementation. I just need to know what these concepts mean and that there is a software that abstracts the low-level implementation to a higher-level function that I can just apply. And, in addition, I want to script those steps together without needing to worry about things like the exact memory allocation in each step.

11 http://retractionwatch.com/2015/11/07/its-official-anil-potti-faked-data-say-feds/

Two programming languages, along with the packages they themselves as well as their users provide, have been fundamental in providing a tool set that allows processing, exploration, and analysis of the amount of data contemporary experiments provide. These are R (Ihaka and R. Gentleman, 1996) and Python (Van Rossum and Drake, 1995), along with packages hosted on CRAN/BioConductor (R. C. Gentleman et al., 2004)[12] and PyPI, respectively. Without those tools, performing analysis like I do to obtain the results shown later would not be possible. Note that there are other tools, but they do not play a similarly important role in contemporary data analysis.

1.7.3 *Reproducible workflows*

Keeping past versions of scripts

Requirements for analyses are going to change, and so will the scripts that were used to generate them. There is nothing that is keeping us from updating and changing them of course, but at times it is required to reproduce the behaviour and outcome of an analysis after that. This is where version control systems (VCS) come in. These are software that will keep track of all previous versions of code (and potentially other files as well) in case they are ever needed again. One of the first to be widely used was CVS (Thomas and Hunt, 2003), later Subversion (Pilato, Collins-Sussman, and Fitzpatrick, 2008) and now mostly Git (Loeliger and McCullough, 2012). A company that has been widely successful in promoting the use of git both in general and academia[13] in particular is GitHub (Dabbish et al., 2012).

Defining workflows with a single entry point

A challenge that often is underestimated is that even when all of the code that was used in order to generate the result is provided, it is often not trivial know which script generates which part of the analysis, and which other scripts, analyses, or data it depends on. Hence it is not only important to provide the code as it is, but also information on how to run it, so to not only have all the bits and pieces but a way to connect them to a workflow from the data to the results. Some tools have been proposed to manage scientific workflows like KNIME (Warr, 2012) and Taverna (Wolstencroft et al., 2013), but they are heavy monolithic pieces of software whose use has never taken off. A simple and lightweight alternative is GNU's `make` (Stallman and

12 including *dplyr* (Wickham and Francois, 2014) and *ggplot2* (Wickham, 2011), used extensively in the analyses I performed

13 Github has become the *de facto* standard in academic code sharing. The technology for version tracking is state of the art, the functionality of their website unparalleled and they offer their services to academics for free: https://education.github.com/

McGrath, 1991; Stallman, McGrath, and P. D. Smith, 2004), which was originally designed for tracking dependencies in software compilation, but has been proposed to enable reproducible scientific workflows (Schwab, Karrenbach, and Claerbout, 2000). It has been extensively used in the analyses because of its single entry point (the *Makefile*) and simple definition of dependency rules.

Generating reports directly from the analysis

While the above tools take care of performing and keeping track of analyses, the crucial point that is missing for reproducible research is integration with reporting. A tool to combine R code with written text and annotation (as opposed to technical documentation) is *knitr* (Xie, 2014) that can work with either Latex or Markdown. For Python, the IPython notebooks (McKinney, 2012) (now JupyterLab[14]) has seen wide adoption (Shen, 2014).

1.8 MOTIVATION AND OUTLOOK

With sequencing technologies becoming more and more commonplace for cancer diagnosis in both the clinical (The Cancer Genome Atlas Research Network et al., 2013; International Cancer Genome Consortium et al., 2010) and preclinical setting (Garnett et al., 2012; Barretina et al., 2012; Iorio, Knijnenburg, et al., 2016), there is a wealth of molecular data available that likely harbours yet undiscovered disease markers and treatment opportunities. Efforts like the TCGA/ICGC have pioneered this characterization on a large scale, offering the opportunity to derive large amounts of information about individual tumours. This represents a step forward over basing the treatment of an individual on the tissue of tumour origin alone, with mutational screens and sensitivity markers providing a first glimpse of the direction that personalized medicine is going to take.

Investigations of cancer drug response and survival are mostly focussed on DNA mutations. This has yielded multiple markers, however, they are rarely put in functional context of the signalling aberrations they create. Because the amount of phosphorylation data available lags behind the one of genomic data, gene expression may be a valuable substitute.

One way gene expression has shown promise is using global clustering to identify subtypes of the disease, and associating those with different drug response or survival outcome. I believe that stratifying patients in a reproducible manner should involve both the overall clustering of the data as well as more specific, functionally relevant readouts. An ex-

14 http://blog.jupyter.org/2016/07/14/jupyter-lab-alpha/

ample of those functional readouts is the expression level of pre-defined pathway sets - mostly using the GSEA scores - that has been reported in numerous publications, but also more sophisticated pathway methods that attempt to quantify the signal flow, such as SPIA (Tarca et al., 2008), PARADIGM (Vaske et al., 2010), or Pathifier (Drier, Sheffer, and Domany, 2013).

These methods, however, stand at odds with the notion that signalling in animal - and thus human - cells is tightly post-translationally regulated, since they are based on mapping the mRNA expression of the signalling proteins, in one way or another. To date, it is largely unexplored how much those pathway-expression-based methods are able to make statements about the signal transmission in a protein network.

This is why, after establishing a baseline for gene set methods to explain drug response (chapter 2), I followed different strategy: use a large body of publicly available perturbation experiments to quantify gene expression responses to a specified set of stimuli (chapter 3), and then use those genes to infer the upstream signal required to change their expression (Bild, Yao, et al., 2005; Gatza, Lucas, et al., 2010; Parikh et al., 2010). This is one way by which–only using gene expression data–I can indirectly observe activity instead of just mRNA expression (chapter 4), as well as apply similar signatures for potentially synergistic drug combinations (chapter 5).

2

GENE SET METHODS FOR DRUG RESPONSE

Pathway methods are often used in a cancer context, both for cell lines and primary tumours. Most of the time, the method of choice is to take a gene set from either Gene Ontology (GO) (Ashburner et al., 2000), KEGG (Kanehisa and Goto, 2000) or Reactome (Croft et al., 2011), and calculate a combined expression score using either a Fisher's exact test (e.g. by a tool called DAVID[1]) if one is to test gene sets against differentially expressed genes, or some variant of Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) if the sets are pre-defined, but one wants to avoid cutting of continuous expression values at an arbitrary threshold. There are, however, more advanced pathway methods available. Signalling Pathway Impact Analysis (Tarca et al., 2008) and Differential Expression Analysis for Pathways (Haynes et al., 2013) take into account the directionality and sign of edges in a pathway. Pathifier (Drier, Sheffer, and Domany, 2013) calculates probable information flow between the set items. PARADIGM (Vaske et al., 2010) employs a Bayesian framework that models translation, activity, and interactions. I will leave the more complex methods for a later chapter and focus on GSEA using different gene sets here.

GSEA using GO gene sets is ubiquitous, often following a differential expression analysis to see which higher-level function the differentially expressed genes mediate. After computing the enrichment scores, our list of genes is condensed down to a list of significantly enriched GO categories that may be related to the phenotype we are observing. This may work very well in some cases. There are, however, a couple of caveats to observe: (1) a gene does not exclusively belong to one process; we might very well get a significant p-value only caused by the overlap between different sets, (2) if we test all categories and correct by false discovery rate we might dilute our signal so much that small categories can no longer be significant, or (3) the process that did indeed cause our phenotype does not correspond to a gene set at all (this can be due to missing biological knowledge, annotation errors, or simply the fact that curators have not yet added a certain gene to a certain category). Maybe the most dangerous caveat of them all is that once we see our list of resulting categories, we are inclined to pick out category that "makes sense". Taking this selection of desired categories on its head, we may also be inclined to overlook a category that we

1 note that although this tool is still widely used, it has last been updated in 2010 and misses a lot of annotations (Wadi et al., 2016)

don't want to see, e.g. because the involved process is already known in literature and we could not publish our new findings in a high-impact journal. The aim of this chapter is to illustrate these issues.

I will use this chapter to examine which processes are involved in making cancer cell lines sensitive or resistant to different drugs in the GDSC panel (Garnett et al., 2012; Iorio, Knijnenburg, et al., 2016). I will not filter the gene sets I use, to see how well represented signalling pathways are among the top hits for drug sensitivity, where they are known to play a pivotal role for targeted therapies (Garnett et al., 2012; Iorio, Knijnenburg, et al., 2016; Yap and Workman, 2012).

Results obtained in section 2.3 contributed the pathway scores for latest publication of the GDSC screening. All analyses, plots, and written text in this thesis I produced myself:

Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, **Schubert M**, Aben N, Gonçalves E, Barthorpe S, Lightfoot H, Cokelaer T, Greninger P, van Dyk E, Chang H, de Silva H, Heyn H, Deng X, Egan RK, Liu Q, Mironenko T, Mitropoulos X, Richardson L, Wang J, Zhang T, Moran S, Sayols S, Soleimani M, Tamborero D, Lopez-Bigas N, Ross-Macdonald P, Esteller M, Gray NS, Haber DA, Stratton MR, Benes CH, Wessels LF, Saez-Rodriguez J, McDermott U, Garnett MJ. "*A Landscape of Pharmacogenomic Interactions in Cancer*". **Cell** (2016).

2.1 METHODS USED THROUGHOUT THIS THESIS

2.1.1 *Gene sets*

To obtain Gene Ontology sets, I used the *BioMart* R package (Smedley et al., 2009) to query the Ensembl (Hubbard et al., 2002; Yates et al., 2016) `hsapiens_gene_ensembl` database for all HGNC symbols that had a Gene Ontology (Ashburner et al., 2000; Gene Ontology Consortium, 2004) ID (`go_id` field) associated with them, yielding three main categories (biological process, molecular function, cellular compartment) with 16413 gene sets covering 18806 genes total[2]. For Reactome (Croft et al., 2011), I downloaded the file *ReactomePathways.gmt*[3]. It contained a total of 1675 pathways covering 7852 genes.

For other gene sets, I used the Enrichr platform (E. Y. Chen et al., 2013) and the gene sets the authors assembled in their GitHub repository[4]. They encompassed gene sets for 35 pathway and pathway-related resources, including Gene Ontology, Reactome (where I queried the original databases to obtain more up-to-date gene lists), as well as

2 query of Ensembl Biomart on March 1st 2016
3 http://www.reactome.org/pages/download-data/
4 https://github.com/yokuyuki/Enrichr

KEGG (Kanehisa and Goto, 2000) (that already used the last non-commercial release).

2.1.2 *Gene Set Variation Analysis (GSVA)*

Gene Set Enrichment Analysis (Subramanian et al., 2005) is the *de facto* standard to compute the expression level of a set of genes. It uses as an input a ranked list of genes (*e.g.* fold changes). It then computes the running sum of a set of interest by starting at the beginning of this list and adding a score if the current gene is in the set, or subtracts a score otherwise. This can be summarised like the following:

$$s_{g+1} = s_g + \begin{cases} 1/n_{set} & g \in set \\ -1/(n - n_{set}) & g \notin set \end{cases}$$

A schema of this calculation is shown in figure 2. In the case of GSEA, the overall score is the maximal deviation from zero. As is shown in the example, this leads to a bimodal distribution of scores when testing different sets or the same set on different samples, because even if the genes in the set of interest are evenly spread, there will always be a deviation. As GSEA is commonly used to compute the significance of enrichment between two conditions (left panels in figure 2), this is not a problem: we can obtain the distribution of scores under the null hypothesis by shuffling the labels of the reference and the samples we are looking at, and then compute the empirical p-value as quantile of this distribution. This, however, also means that we need to compare two conditions in order to do this reliably. We can not compute enrichment scores for each individual sample.

Gene Set Variation Analysis (Hänzelmann, Castelo, and Guinney, 2013) solves this: instead of taking the maximal deviation, it takes the difference between maximum positive and negative enrichment score. This directly yields a unimodal distribution of enrichment scores in different samples that can hence be used in statistical tests that assume normality[5]. As I am interested in correlating one continuous value (drug sensitivity) with the set enrichment score, using GSVA (and the *GSVA* R package) instead of GSEA is the natural choice.

2.1.3 *Drug associations using the half-maximum inhibitory concentration (IC50)*

The original GDSC data set contained different dilutions of drugs that the cell lines in the panel were subjected to, measuring how much it

5 raw gene enrichment scores could still be used in nonparametric tests, but they are usually less powerful

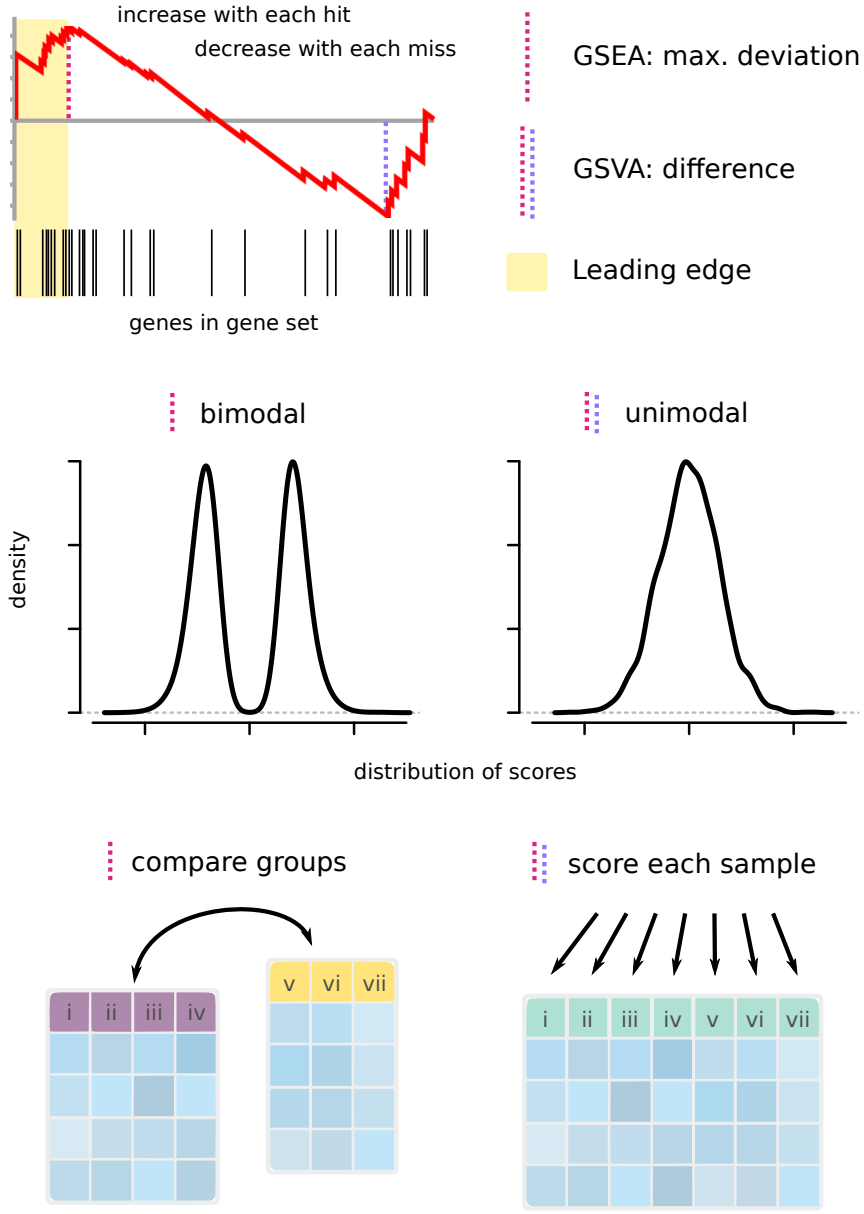

Figure 2: Calculation of the running sum statistic for GSEA and GSVA. Calculation of the running sum statistic (top left) and absolute deviation for enrichment score in case of GSEA vs. the difference in GSVA. Genes are ordered by differential expression, genes that are in the query set are indicated by black bars. The red line indicates the running sum score where a score is added each time there is a hit and subtracted otherwise. GSEA hence produces a bimodal distribution of scores (left), while GSVA produces a unimodal distribution (right). This is why the former needs label shuffling of two conditions (bottom left) to compute empirical p-values, while the latter produces scores for each sample (but no statistical significance; bottom right).

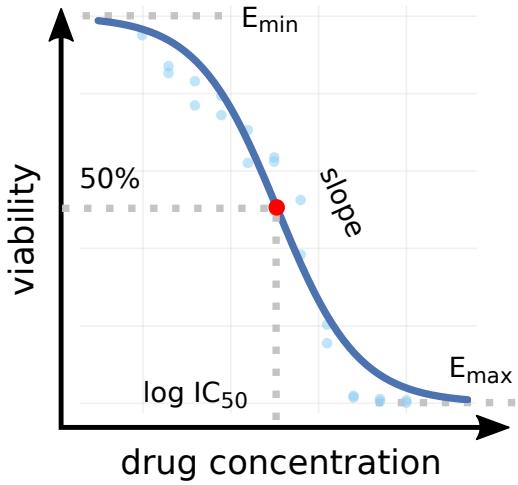

Figure 3: Calculation of IC_{50} values.Calculation of IC_{50} values. Cell viability is measured at different drug concentrations and then a drug response curve is fitted to the data. The halfway point of viability between the minimum (E_{min}) and maximum effect (E_{max}) is the IC_{50}. The curve is defined by these three values and the steepness of the slope.

interfered with their growth. Since I have got a pathway score for each cell line, I also need a single value corresponding to the sensitivity to a given drug. One way to do this is to measure the growth inhibition at different concentrations, and then fit a dose-response curve to the data points, interpolating (or extrapolating, if necessary) the concentration at which the half-maximal inhibition occurred. This term is referred to the IC_{50} value, and has already been calculated in (Iorio, Knijnenburg, et al., 2016). The curve to fit is of sigmoid shape (figure 3) and has the formula:

$$x = IC_{50}\left(\frac{y - E_{min}}{E_{max} - y}\right)^{-slope}$$

I obtained already processed gene expression matrix from the GDSC cell lines and their fitted IC_{50} values to 265 public drugs from the GDSC publication (Iorio, Knijnenburg, et al., 2016). I performed a linear regression using the `lm` function in R between the gene set score as an independent variable (S_j, where j corresponds to each different phenotype from 1 to k; this could *e.g.* be pathways or the presence of a mutation) and the log_{10} of the $IC50$ in micro-molar as the response variable (D_i, where i is the drug index). I regressed out the contribution of individual tissues by including it as a covariate (T) in the fit.

$$D_i \sim T + S_j \qquad \forall i \in drugs\ \forall j \in phenotypes$$

In other words, for each drug D_i, I fit the following model for all cell lines c in the GDSC panel.

$$\begin{matrix} D_i^{c_1} \\ D_i^{c_2} \\ D_i^{c_3} \\ \vdots \end{matrix} \sim \begin{matrix} T^{c_1} \\ T^{c_2} \\ T^{c_3} \\ \vdots \end{matrix} + \begin{matrix} S_j^{c_1} \\ S_j^{c_2} \\ S_j^{c_3} \\ \vdots \end{matrix}$$

I performed this association between every drug and all gene set scores, yielding an effect size (how many units of drug response changed per unit of enrichment score) and p-value for each pair. I corrected the p-values for each pair using the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). In addition, I performed these associations using each tissue separately:

$$D_i \sim S_j \qquad \forall i \in \textit{drugs}\ \forall j \in \textit{phenotypes} \mid T^c = t; \quad \forall t \in \textit{tissues}$$

In this case, I only include cell lines c whose tissue Tequals t and build models for each tissue separately.

$$\begin{matrix} D_i^{c_1} \\ D_i^{c_2} \\ D_i^{c_3} \\ \vdots \end{matrix} \sim \begin{matrix} S_j^{c_1} \\ S_j^{c_2} \\ S_j^{c_3} \\ \vdots \end{matrix}$$

2.2 CELL LINE DRUG RESPONSE

2.2.1 *Associations with Mutations*

Associations between drug response and mutated genes have already been published with the 2012 and 2016 versions of the GDSC screening (Garnett et al., 2012; Iorio, Knijnenburg, et al., 2016). I reproduce them here in order to ensure that the associations I obtain are the same as the ones previously published. P-values vary slightly between the two because the cell lines included in this study are not exactly the same as in the original article. But the overall results (volcano plot in figure 4 and associations in appendix A.1) very much agree: the strongest hit in both cases is that *TP*53 mutations correlate with resistance to Nutlin-3a, drugs that specifically target mutant *BRAF* require such a mutation to be effective (Dabrafenib, PLX4720), and MEK inhibitors work better with mutations in *KRAS* or *NRAS*.[6]

6 The pan-cancer volcano plot has been removed in the published version, but is available here: http://www.cancerrxgene.org/gdsc1000

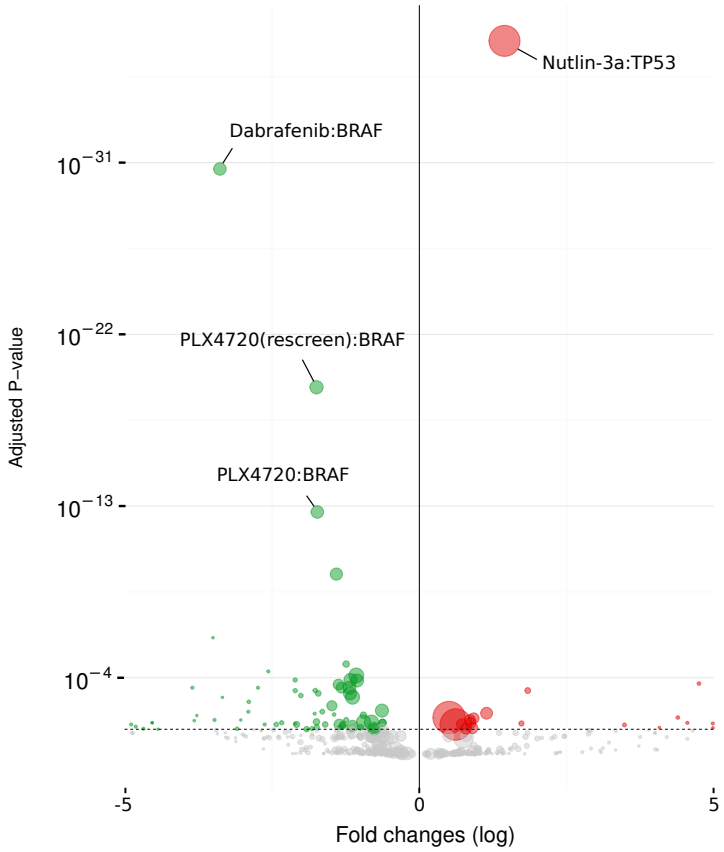

Figure 4: Volcano plot of associations between driver mutations and drug response. Effect size is fold changes between cell lines harbouring a mutated vs. a wild-type copy on the horizontal axis, FDR-adjusted p-values on the vertical axis.

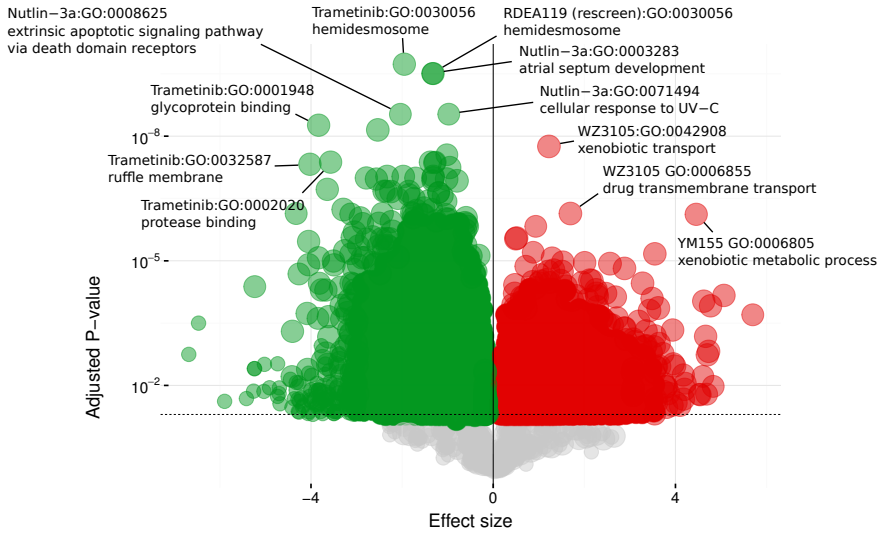

Figure 5: Volcano plot of associations between expression of Gene Ontology categories and drug response. Effect size is standard deviations of the score on the horizontal axis, FDR-adjusted p-values on the vertical axis.

2.2.2 *Associations with Gene Ontology categories*

The most significant associations for Gene Ontology (Croft et al., 2011) are much less clear than the ones for mutations. I used all categories for "biological processes" and "molecular function" between 5 and 500 genes to calculate gene set scores using GSVA. The results for their correlations with drug response are shown in figure 5 (associations in appendix A.1).

Among the the top drugs, there are MEK inhibitors (Trametinib, RDEA119), p53-stabiliser Nutlin-3a, and the multi-kinase inhibitor WZ3105. The biological processes that they are involved in have no obvious connection with their mechanism of action: While the "extrinsic apoptotic signaling pathway via death domain receptors" and "cellular response to UV-C" could somehow be linked to Nutlin-3a via p53-mediated apoptosis and DNA damage respectively, there is no obvious connection between glycoprotein binding, protease binding, the ruffle membrane, or hemidesmosomes and MEK inhibitors. Similarly, "xenobiotic metabolic process" gives a hint that YM155 may be inactivated by modification of the drug, but it does not tell us anything about the mechanism of the drug (it binds the promoter of Survivin, suppressing its expression[7]) or its possible indications. Overall, there are so many significant associations that it is necessary to select interesting categories either before or after computing those in order to be able to interpret them.

7 https://www.caymanchem.com/product/11490

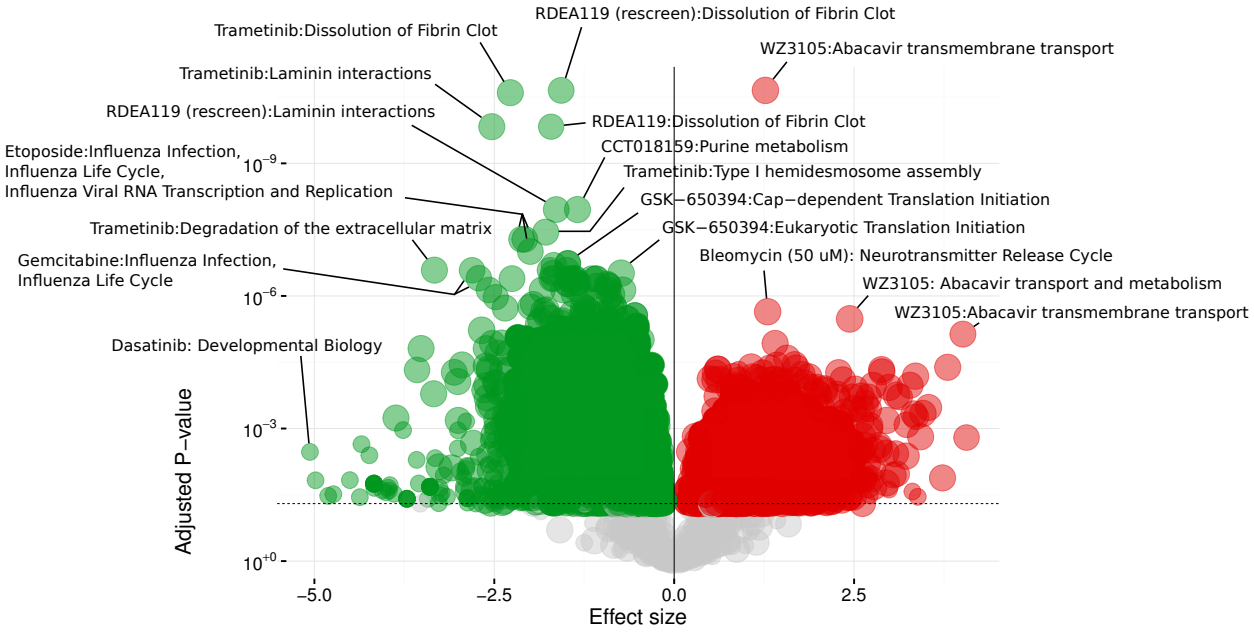

Figure 6: Volcano plot of associations between expression of Reactome pathways and drug response. Effect size is standard deviations of the score on the horizontal axis, FDR-adjusted p-values on the vertical axis.

2.2.3 *Associations with Reactome pathways*

Compared to Gene Ontology, the drug associations with Reactome pathway enrichment using GSVA are maybe even harder to interpret (figure 6 and associations in appendix A.1). The MEK inhibitors Trametinib and RDEA119 associate most strongly with fibrin clot dissolution and laminin interactions (a fibrous protein present in the basal lamina of the epithelia). Gemcitabine and Etoposide are more effective if Influenza-related pathways are expressed, possibly hinting at involvement of the DNA replication machinery in promoting sensitivity to a Topoisomerase inhibitor and nucleoside analogue, respectively. Resistance to WZ3105 is again associated with a process acting on the drug, but this time it is export rather than modification. There is no obvious link between Bleomycin (which induces DNA double strand breaks) and the "Neurotransmitter Release Cycle". Again, there are so many significant associations that we need to limit the pathways in order to make sense of them.

2.3 PATHWAY-RESPONSIVE GENES: THE SPEED PLATFORM

Another possibility is to start off with fewer gene sets that we know we are interested in. In the case of cancer signalling and drug response, these could be the signalling pathways that we know are involved.

The original SPEED platform (Parikh et al., 2010) consists of signatures of 11 pathways, derived and comprised of the genes responsive to a total of 215 experiments (the perturbations between a control and a perturbed condition). They use a consensus gene signature across multiple experiments, perturbing agents, and other conditions to arrive at a gene list that corresponds to pathway activation in a wide range of conditions. These consensus signatures of pathway perturbations are distinct from the expression level of pathway members that I described above, as they are a downstream readout and not the expression status of the signalling molecules.

2.3.1 *Separability-optimised Gene Sets*

The authors of the original SPEED publication used four parameters to generate gene lists from their input experiments:

- Z-score cutoff: the top n% of upregulated genes
- Total expression cutoff: the top m% of genes considering their basal expression in each experiment
- Experiment overlap: the percentage of experiments for which the other two conditions must be met
- Uniqueness: whether only genes should be returned that were unique to the stimulation of a specific pathway

With the SQLite database[8] and Python query tools the authors provided, I extracted gene lists using the above parameters. First, I used the default parameters in their implementation, which was to include all genes that were top 5% of up-regulated genes by z-score, overall top 50% of expressed genes, in at least 20% of the experiments per pathway, and disregarding whether the gene was in any other pathway or not. Using these default parameters, I obtained scores that were highly correlated between the different pathways, as shown in figure 7 (left).

To counteract this problem, I extracted gene lists for different combinations of the four parameters:

- Z-score: 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 9, 10, 11, 12, 15, 20, 25
- Total expression: 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
- Overlap: 5, 10, 20, 30, 40, 50, 60, 70, 80
- Uniqueness: True or False

8 http://speed.sys-bio.net/SPEED_db.zip

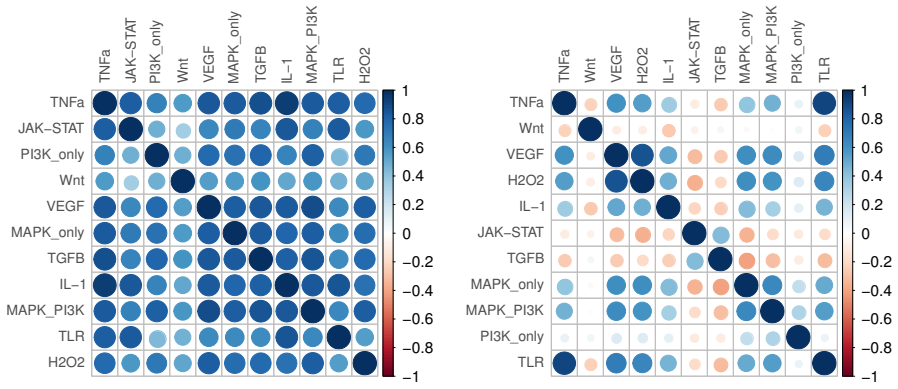

Figure 7: Correlation plots for GSEA scores per pathway across cell lines for the original SPEED lists (left) and my separation-optimized version (right).

For each combination, I optimised the order obtained by GSEA scores between control and stimulated experiments. For each combination of parameters, *i.e.* for each of the 4950 gene lists, I calculated the raw enrichment score for each pathway and cell line, yielding a total of 55 million enrichment scores (as the ordering is non-parametric, the raw GSEA score is good to use here while being quicker to compute). I then went on to, for each pathway, construct a precision-recall curve that quantified how well the GSEA scores were able assign lower pathway activation scores to the control arrays than the perturbed arrays (where, in the case of an inhibition I performed GSEA using negative z-scores).

The set of control arrays comprised all the un-stimulated arrays in the database, and the stimulated set of all unperturbed arrays in the database where a certain pathway was perturbed. I used as a measure of performance the area under the precision-recall curve. A perfect ordering (that is, all control arrays and then all stimulated arrays) corresponded to a precision-recall AUC (prAUC) of 1, while a random ordering would respond to a prAUC about 0.5.

By using not only the matched control arrays to the perturbed arrays but all arrays present in the data set, I allow in my resulting signature the cross-activation of pathways while minimizing the fit to random differences in gene expression by different initial conditions.

I split the data set (both control and perturbed arrays) in five different subsets, where four of the five were the designated training set and the fifth the test set. I calculated the prAUC for all the parameters described in the previous section, and chose the set with the highest score. I then went on to the part that was not used in training and quantified the prAUC there as well. I performed the whole process five times, with another subset functioning as the test set each time. I then chose the set that the highest prAUC in the test set, or in the training set if it was lower than in the test set. I did not simply select the

Table 2: Parameter selection overview for each pathway. Z: Z-value cutoff (0.25-20%), O: array overlap (20-80%); E: overall expression cutoff (5-100%), U: unique (+) or all (-) genes considered. Gene lists had to be between 50 and 250 genes to be considered. Values for optimized and original lists are shown. Precision-recall AUC shown for training, cross-training for optimized lists, and whole data set for optimized lists vs. the one obtained by using the original lists.

2*pathway	cutoffs used				# genes in list		precision-recall AUC		all arrays	
	Z	O	E	U	opt	original	train	cv	opt	original
H2O2	20	50	20	+	191	60	0.99	1.00	**0.99**	0.66
IL-1	20	60	20	+	75	141	0.91	0.99	**0.93**	0.85
JAK-STAT	20	20	70	+	162	114	0.78	0.87	**0.81**	0.68
MAPK_only	20	70	10	-	65	559	0.94	0.97	**0.95**	0.46
MAPK_PI3K	15	20	5	-	171	118	0.82	0.89	**0.84**	0.63
TLR	6.5	40	50	+	78	181	0.88	0.91	**0.89**	0.81
PI3K_only	15	20	30	+	227	67	0.80	0.97	**0.83**	0.49
TGFB	20	20	60	+	119	142	0.78	0.88	**0.80**	0.68
TNFa	1.5	50	70	-	56	259	0.77	0.99	**0.81**	0.66
VEGF	12	20	30	+	121	56	0.92	0.94	**0.92**	0.84
Wnt	7.5	30	5	+	195	83	0.93	0.94	**0.93**	0.65

highest highest AUC in the test set because I would not want to select a model that performed badly on the training set to begin with. This selection procedure can be represented using the following formula:

$$selected = max^{all\ runs}(min^{per\ run}(prAUC^{train}, prAUC^{test}))$$

For the optimised lists, I observed a much lower overall correlation of the pathway scores (figure 7, right). The values for the different gene list cutoff parameters that I selected after optimisation are listed in table 2, including the number of genes in the signature and prAUC (training and test set) compared between the original cutoffs used in the query tool and my selection. For all the pathways, the optimisation of parameters yielded a better separation of control- vs. perturbed arrays.

2.3.2 *Associations between Pathway Scores and Tissues*

As a control, I compared the inferred pathway activation scores between different tissues. If the assigned scores per tissue are biologically meaningful, I would expect to find well-established literature evidence supporting them. An overview heatmap of pathway activation scores is shown in figure 8. Relating this to known biology, the scores seem well supported by previously known evidence. A couple of examples to illustrate this are listed below.

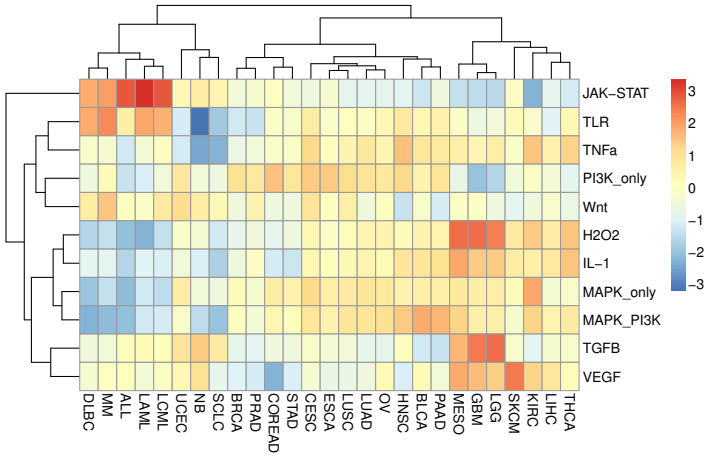

Figure 8: Heatmap of inferred pathway activation scores for different tissues. Pathways in rows, TCGA tissue labels in columns, relative pathway activation indicated by colour. Rows and columns are clustered so that similar tissues and pathways are shown close to each other, with branch lengths indicating distance.

- JAK-STAT signalling is well-known to be upregulated in blood cancer cells (Dutta and W. X. Li, 2013; Vainchenker and Constantinescu, 2012), including in particular AML (H. J. Lee et al., 2012; Danial and Rothman, 2000), CML (Danial and Rothman, 2000), ALL (Vainchenker and Constantinescu, 2012), and DLBC (Gupta et al., 2012). This correlates well with the inferred activity scores for JAK-STAT signalling. Similarly, TLR signalling has been shown to be induced by CpG oxydinucleotides in B-(DLBC, MM) and dendritic myeloid, but not T-cells (Rothenfusser et al., 2002).

- High production of reactive oxygen species has been shown to occur in certain mesotheliomas (Kahlos et al., 1999), and gliomas (Drukala et al., 2010).

- TGFB (Yamada et al., 1995; Kjellman et al., 2000) and VEGF (D. A. Reardon et al., 2008) expression have been shown to be increased in gliomas, correlating with malignancy (Kjellman et al., 2000; Leon, Folkerth, and Black, 1996). Both were also increased in mesothelioma (Kuwahara et al., 2001; Aoe et al., 2006), and VEGF in melanoma (Rajabi et al., 2012; Gajanin et al., 2010) as well.

- MAPK_only/MAPK_PI3K signalling seems to be evenly distributed among non-blood cancer tissues. We found the highest activity of MAPK and MAPK_PI3K can be found in KIRC and BLCA/PAAD, respectively. However, the difference was a lot

less apparent than for other pathways, suggesting that MAPK and EGFR pathway are important for all solid cancers.

2.3.3 *Pan-Cancer Drug Associations*

I performed a linear regression between the obtained pathway scores and the IC_{50}s for all cell lines and pathways while correcting for tissue labels (used as covariate; details section 2.1.3). I adjusted p-values by controlling the false discovery rate and visualized the result as volcano plots and individual fits.

Linear associations of inferred pathway activity within all tissues and drug response are shown in figure 9 (associations in appendix A.2). Negative regression slopes (left side of the graph, green) indicate sensitivity markers, *i.e.* a higher pathway activation score correlates with a lower IC_{50}. Positive regression slopes (right side, red) indicate resistance markers, *i.e.* higher activation scores correlate with higher IC_{50}s.

Sensitivity markers:

- RDEA119 (Iverson et al., 2009), PD-0325901 (Ciuffreda et al., 2009), and CI-1040 (Allen, Sebolt-Leopold, and Meyer, 2003) are all MEK inhibitors and are thus to be expected to be more effective in cell lines where MAPK signalling is more active. In fact, the strongest associations are between those drugs and MAPK_PI3K signalling. However, MAPK and PI3K are difficult to distinguish in expression response due to pathway crosstalk (Parikh et al., 2010).

- BIBW2992 and Gefitinib showed higher efficacy with PI3K_only activity. As both are are EGFR inhibitors and PI3K is known to cause resistance to those (Jeannot et al., 2014), this result is surprising because the association is stronger than with MAPK_only or MAPK_PI3K. It may be because the authors of the SPEED platform chose to include MEK inhibition as condition for PI3K activation.

- I found sensitivity correlating with Wnt activity for the drugs Etoposide, QS11, and GSK-650394. QS11 modulates ARF activity and β-catenin localisation (Q. Zhang et al., 2007), which may offer a treatment strategy for Wnt-driven tumours. GSK-650394 targets SGK1, which is activated by Wnt/β-catenin signalling and has been shown to inhibit ROS-induced apoptosis in liver cells (Tao et al., 2013). Etoposide induces DNA damage and senescence, where this process may be inhibited by negative feedback by SFRP1 (Elzi et al., 2012) due to Wnt signalling.

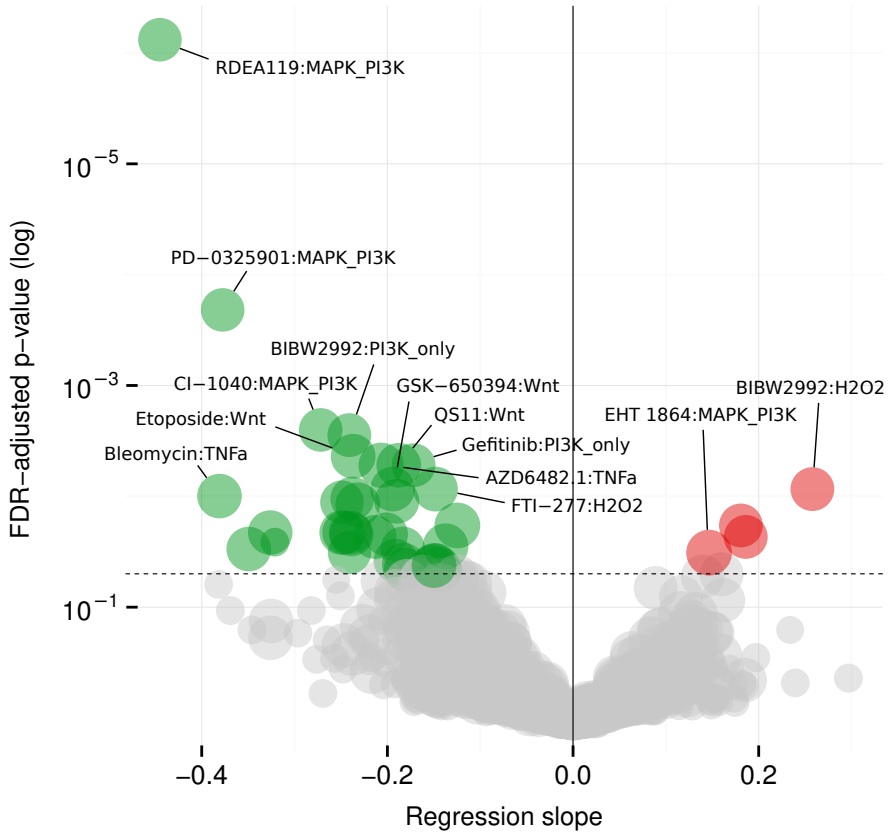

Figure 9: Volcano plot of linear associations of inferred pathway activity within all tissues and drug response. Tissue of origin used as a covariate in the regression. P-values FDR-adjusted. Negative regression slopes (left side of the graph, green) indicate sensitivity markers, *i.e.* a higher pathway activation score correlates with a lower IC_{50}. Positive regression slopes (right side, red) indicate resistance markers, *i.e.* higher activation scores correlate with higher IC_{50}s.

- FTI-277 is more effective when reactive oxygen response (H2O2) is active. Farnesyl transferase inhibitors are known to induce DNA damage via ROS (Pan et al., 2005), which may cause growth arrest or apoptosis in cells that already suffer ROS damage.

- AZD6482 is a PI3K inhibitor (Nylander et al., 2012), for which cells show increased sensitivity in our study when TNFa signalling is active. While the latter can be both oncogenic and tumour-suppressive (Pikarsky and Ben-Neriah, 2006), it has been shown that PI3K activation is necessary for NFKb-mediated cell survival in DLBC (Kloo et al., 2011) and the combination of PI3K inhibition and active TNFa is known to cause apoptosis in vitiligous keratinocytes (N.-H. Kim et al., 2007).

- TNFα (Sleijfer et al., 1998) signalling has been shown to be increased after Bleomycin treatment, thereby mediating cytotoxicity. It can be hypothesised that if this pathway is active in cell lines, they are more likely to be affected by this.

Resistance markers:

- EHT 1864 is a Rac-family GTPase inhibitor (Shutes et al., 2007), and MAPK_PI3K signalling is associated with resistance to this drug. Rac1 is known to be involved in MAPK signalling specifically for cancer development (Khosravi-Far et al., 1995). Hence, cells with a higher MAPK activity may be less susceptible to Rac1 inhibition.

2.4 DISCUSSION

2.4.1 *Cell line drug response*

The importance of mutations, especially when they are drivers, and their role in a cell line's response to the different drugs is well established. With the new GDSC release (Iorio, Knijnenburg, et al., 2016) the authors uncovered previously unknown links that may ultimately lead to new clinical indications, or prioritise the development of certain drugs over others.

In contrast to this, the biological meaning of the top associations between gene set or pathway scores and drug response I obtained is doubtful at best. What does it mean to have an association of the "T cell receptor" in Head and Neck Squamous Carcinoma (HNSC)? There are no immune cells involved in the culturing of HNSC cell lines, for example.

Hence, we have two options: we either need to link seemingly unrelated gene sets back to the process that actually caused a difference in

drug response by looking for evidence that may support it, or we need to pre-select gene sets that may be relevant for drug response.

For the first case, we can not easily follow the chain of causality between a biological process that mediates differential drug response and its downstream readout as change of gene expression. The second alternative does not solve this, but leads to a much higher probability to catch causative gene sets when we already select candidates of exactly this by prior knowledge.

2.4.2 *Pathway-responsive genes*

Using pathway-responsive genes instead of pathway expression to infer signalling activity makes a lot of sense: we look at the footprint of the actual signalling activity (the expression changes downstream of a signalling pathway) and not the potential mediators (protein kinases, among others) by means of mRNA expression level that is a lot further removed from the actual signalling going on in a cell. However, the SPEED platform has some issues in terms of the level in which its pathways correlated to one another: If we did calculate drug associations with the original scores, hits would be groups where a given drug is correlated with all pathways in about the same extent. This is likely due to how they only evaluated gene lists by their overlap with Gene Ontology categories, and not how well its enrichment scores are able to differentiate between microarrays where a given pathway perturbation is present and those where it is absent. The original version was hence not very useful for my purposes.

There is a need for scores that are more potent in distinguishing the gene expression footprints from one pathway to another. A way to do this in the existing platform is what I did: no longer require that the signature genes from all pathways are obtained using the same cutoffs of the author's 4 parameters, but instead optimise them in way so that they are best able to tell apart the pathways one from another. Looking at the correlations between the original cutoffs and the ones I suggest after the cross-validated optimisation, I know that this worked at least on those terms.

The next question to answer is whether we get more meaningful drug associations. Looking at the volcano plot, associations are, for one, not between a drug and all pathways, and for the other much better supported when searching for literature corresponding to the top hits.

However, there is still a number of potential issues with the current model, many of which can be traced back to using the original platform:

- The authors use raw microarray data as well as processed data. For the processed data, we have no idea of what the original authors did with their data to arrive at the expression levels they

report. Sometimes they report this in their respective experimental or data analysis procedures, but often they don't. There is a potential of a variety of biases that we can not control for.

- The platform as it currently stands needs four parameters to be specified, each of which corresponds to a somewhat arbitrary cutoff. Even if we ignore this, they limit the number of signature genes in a way that does not support down-regulated genes at all (as the z-scores are filtered by top percentile only).
- We bias the selection of genes to the ones most commonly found in microarrays. If a gene is highly upregulated but not present in arrays and thus failing the overlap cutoff, we would lose it.
- MAPK inhibition was in the curated set of PI3K activators. This is an error in curation and could explain many PI3K associations.

Hence I argue that, while the optimisation of parameters yielded vast improvements in terms of correlatedness of scores and resulting drug associations, it still has enough drawbacks to suggest that a different approach keeping the overall idea of using pathway-responsive genes as signature for pathway activity may be worth exploring. I describe the approach I developed in the next chapter.

3

BUILDING AN IMPROVED MODEL OF PERTURBATION-RESPONSE GENES

There are numerous examples of using Gene Set Enrichment Analysis (GSEA) for a set of signalling molecules in order to infer signalling activity. Most of them, however, use the expression of those proteins directly as a proxy for their activity. My hypothesis is that using this footprint of signalling activity is a better proxy than assuming that it is proportional to the expression of the proteins involved in the signal transduction pathway, which assumes the latter accurately represents not only the protein levels but also their activity. To my knowledge, the only similar efforts are the SPEED database (Parikh et al., 2010) and two studies MCF10A breast signatures (Bild, Yao, et al., 2005; Bild, Potti, and Nevins, 2006; Gatza, Lucas, et al., 2010; Gatza, Kung, et al., 2011; Gatza, Silva, et al., 2014).

Whereas the former used a smaller but still significant set of publicly available experiments, the only analysis they did using their signatures was to compute the overlap between genes present in their signatures and Gene Ontology categories, and genes between their pathways.

In the latter article the authors experimentally derived pathway-response signatures for a single condition (MCF10A cell line, one perturbation/time point per experiment). Without claiming a direct connection, one of the involved authors later had his medical license revoked and multiple articles retracted because of irreproducibility of results, as investigated by (Baggerly and Coombes, 2009).

A later generation of signatures (Gatza, Lucas, et al., 2010) still had this author involved, but the signatures themselves seem to be well documented and the raw data was deposited in a public microarray repository. Finally, in 2014 the main author (Gatza, Silva, et al., 2014) linked the inferred pathway activity to drug response in a handful of breast cancer cell lines. The MCF10A cell line itself is non-cancerous, so it will represent the behaviour of cells without cancer-causing molecular alterations. While the data in general seems to be of high quality, I am not confident to make any statements about the genes responsive in this cell line can be extrapolated to other tissues, especially because it has previously been shown that signatures generated in a certain lab can correlate more with the lab they were produced in than the biological condition (Chibon, 2013), which I can not control for if there is only one lab producing them.

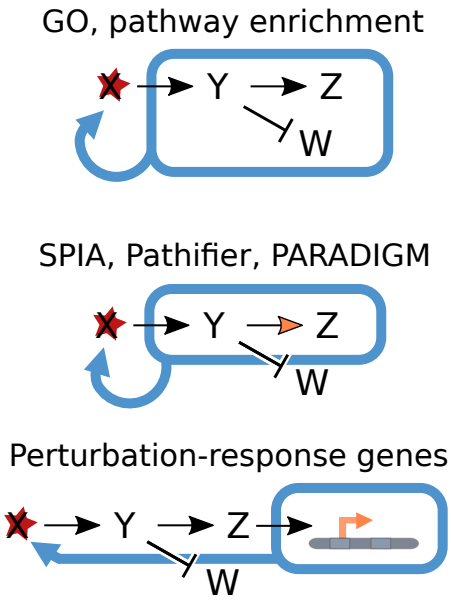

Figure 10: Reasoning about pathway activation. Most pathway approaches make use of either the set (top panel) or graph (middle panel) of signalling molecules to make statements about a possible activation, while our approach considered the genes affected by perturbing them.

So far, it is unknown how well it is possible to derive consensus signatures that provide meaningful estimates of pathway activity across different conditions. Only SPEED computed them, and the authors did not do a detailed investigation of either this, or any large-scale functional association with the pathway scores they obtained. It also relies on a framework that applies multiple arbitrary cutoffs in order to obtain signature sets, even though a simple linear model would abrogate the requirement for such cutoffs, while being able to take into account both up- and downregulated genes.

The chapter represents the model building stage of improved pathway signatures. All analyses, plots, and written text in this thesis I produced myself, while incorporating comments from the coauthors:

Schubert M, Klinger B, Klünemann M, Garnett MJ, Blüthgen N, Saez-Rodriguez J. "*Perturbation-response genes reveal signaling footprints in cancer gene expression.*" **bioRxiv** 065672 (2016). doi:10.1101/065672

3.1 CURATING AND ASSEMBLING A DATABASE OF PUBLICLY AVAILABLE PERTURBATION EXPERIMENTS

3.1.1 *Defining scope and format*

Because the experiments in the SPEED database were curated in a first iteration in 2009 and other pathways were added in 2012, it is

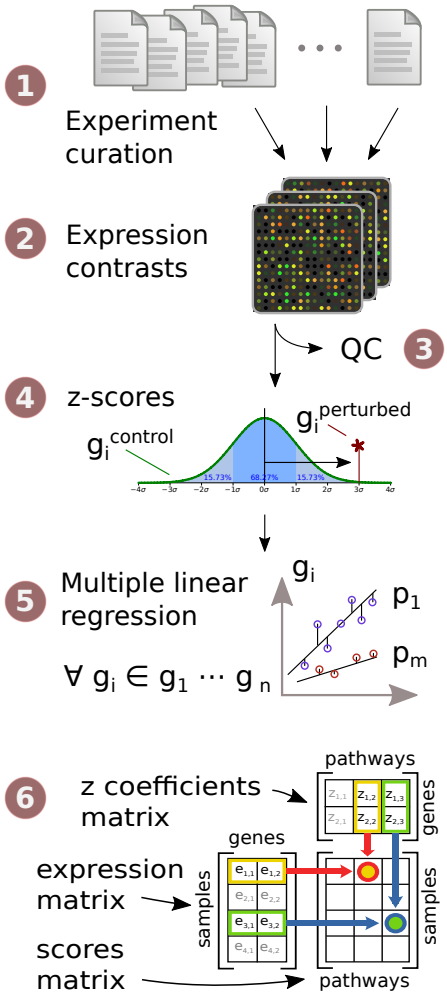

Figure 11: Workflow of data curation and model building. (1) Finding and curation of 208 publicly available experiment series in the ArrayExpress database, (2) Extracting 556 perturbation experiments from series' raw data, (3) Performing QC metrics and discarding failures, (4) Computing z-scores per experiment, (5) Using a multiple linear regression to fit genes responsive to all pathways simultaneously obtaining the z-coefficients matrix, (6) Assigning pathway scores using the coefficients matrix and basal expression data.

quite likely that there are many more relevant experiments that were deposited in ArrayExpress or GEO after the data was curated. Also, previous efforts made no distinction based on the availability of raw data or how many arrays were indeed available in a given contrast. For the latter, I set the requirement to have at least two arrays in the control condition to be able to obtain an estimate of the variability of the expressed genes without a pathway perturbation, and one perturbed array to compare to the controls. I kept the requirement of the original SPEED for the perturbation to last 24 hours or less in order to focus the analysis on primary pathway responses and not long-term rewiring.

I decided on focussing on the pathway set of the original SPEED publication, with the following modifications:

- Focus on EGFR-related signatures instead of MAPK_PI3K, which could be either any of EGFR, MAPK, or PI3K where in the latter two the respective other was inhibited
- MAPK inhibition should be interpreted as inhibition of MAPK, not activation of PI3K (which may happen and we do indeed see this in a couple of cases, but this is not generally true)
- Remove TLR, instead use TNFa and NFkB

Taking into account these modifications, the final pathway sets I re-curated were: EGFR, H2O2, JAK-STAT, MAPK, NFkB, p53, PI3K, Trail, VEGF, and TGFb.

Instead of creating an SQLite database (original SPEED) or flat text files with semi-defined fields (additional pathways for the original platform), I used the markup language YAML[1] (Ben-Kiki, Evans, and Ingerson, 2005) to create index files for all the curated experiments. This has the advantage over a simple tabular format in that it supports structured data, including field names and comments. It has the advantage over XML that it is easily machine- and human-readable. It can hence easily be reused by other projects for which a curation of perturbation-response experiments on the resolution of pathways is of interest. An overview of the format structure is shown in table 3.

3.1.2 *Finding suitable Experiments*

I used the following searches on the ArrayExpress database to find experiments:

- EGFR: (egf* OR "growth factor" OR TGF*alpha OR epiregulin OR heregulin OR neuregulin OR epigen OR betacellulin OR amphiregulin OR iressa OR gefitinib OR cetuximab OR erlotinib

1 short for Yet Another Markup Language

Table 3: Structured data in the YAML file format that I used from experiment curation for different pathway perturbation experiments in ArrayExpress

Identifier	Description
− − −	marker for the beginning of a record
id	the experiment identifier; consists of
accession	the experiment accession in ArrayExpress, e.g. E-GEOD-15986
platform	the microarray platform
pathway	the pathway identifier
cells	a description of the cells used in the experiment
treatment	a description of what the cell were treated with
effect	if the effect was “activating” or “inhibiting”
hours	how many hours the perturbation lasted
control	a list of array identifiers used for the basal condition
perturbed	a list of array identifiers used for the perturbed condition
...	marker for the end of file

OR lapatinib OR AG1478 OR trastuzumab OR herceptin) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:”Transcription profiling”

- H2O2: (h2o2 OR *peroxide) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:”Transcription profiling”
- Hypoxia: (hypoxia* OR HIF1 OR HIF-1) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:”Transcription profiling”
- JAK-STAT: (ifn* OR interferon OR JAK OR STAT OR prolaktin OR EPO OR IL-2 OR IL2 OR IL-3 OR IL3 OR IL-6 OR IL6 OR IL-10 OR IL10 OR IL-13 OR IL13) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:”Transcription profiling”
- MAPK: (b-raf OR braf OR raf* OR mek* OR erk* OR PD*0325901 OR CI*1040 OR PLX* OR U0126 OR AZD6244 OR GSK1120212 OR *metinib OR PD*325901) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:”Transcription profiling”
- NFkB: (LPS OR sa:lps OR tlr OR sa:lipopolysaccharide OR pam3c* OR zymosan OR parthenolide OR sa:11-7082 OR CAY10512 OR sa:*salicyl*) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:”Transcription profiling”
- p53: (sa:*p53 OR sa:mdm2 OR sa:nutlin* OR sa:*radiation OR sa:*radiated OR sa:”dna damage” OR sa:ddr OR sa:*rubicin OR sa:*platin OR sa:docetaxel OR sa:gamma) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:”Transcription profiling”
- PI3K: (SF1126 OR LY294002 OR TGR*1202 OR SAR2454* OR GSK1059615 OR 80*6946 OR Perifosine OR Idelalisib OR PI3K

OR PIK3* OR PTEN OR BEZ235 OR RP6530 OR GDC*941 OR INK1117) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:"Transcription profiling"

- TNFa: (tnf* OR "necrosis factor") organism:sapiens samplecount:[4 TO 10000] raw:true exptype:"Transcription profiling"

- Trail: (rtrail OR trail* OR fas OR fasl) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:"Transcription profiling" array:(A-AFFY-33 OR A-AFFY-44 OR A-AFFY-37 OR A-AFFY-141)

- Trail (alternative): (Genasense OR ABT-737 OR ABT-199 OR obatoclax OR GX15-070 OR casp8 OR caspase) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:"Transcription profiling"

- VEGF: (VEGF* OR sunitinib OR nexavar OR sutent OR pazopanib OR everolimus OR afinitor OR votrient OR Bevacizumab OR sorafenib OR itraconazole OR PDGF* OR PGF OR FIGF OR Withaferin) organism:sapiens samplecount:[4 TO 10000] raw:true exptype:"Transcription profiling"

- TGFb: (tgf OR "tgf-*" OR "bmp*" OR "SMAD*" "transforming growth factor" OR "A 83-01" OR "A-83-01" OR "A83-01" OR "D 4476" OR "D4476" OR "D-4476" OR "GW 788388" OR "GW788388" OR "GW-788388" OR "LY 364947" OR "LY-364947" OR "LY364947" OR "R 268712" OR "R-268712" OR "R268712" OR RepSox OR "SB 431542" OR "SB-431542" OR "SB431542" OR "SB 505124" OR "SB-505124" OR "SB505124" OR "SB 525334" OR "SB-525334" OR "SB525334" OR "SD 208" OR "SD208" OR "SD-208") organism:sapiens samplecount:[4 TO 10000] raw:true exptype:"Transcription profiling"

Those searches yielded between 30 and 650 results in the database, the majority of which had nothing to do with the perturbation I was looking for but the search term was rather mentioned in the experiment description. I read the description, sample attributes, metadata files, and if required the associated article to fill in the required information about each experiment (listed in table 3).

I found a total of 208 submissions encompassing 11 pathways, 572 experiments, and a total of 2687 arrays. This is the same number of pathways in the SPEED platform (Parikh et al., 2010), but about three times as many experiments and more than four times as many microarrays than were used to derived their signatures. Because the collection of (Gatza, Lucas, et al., 2010) is breast cancer only and only consists of a single experiment per signature, my number of pathway

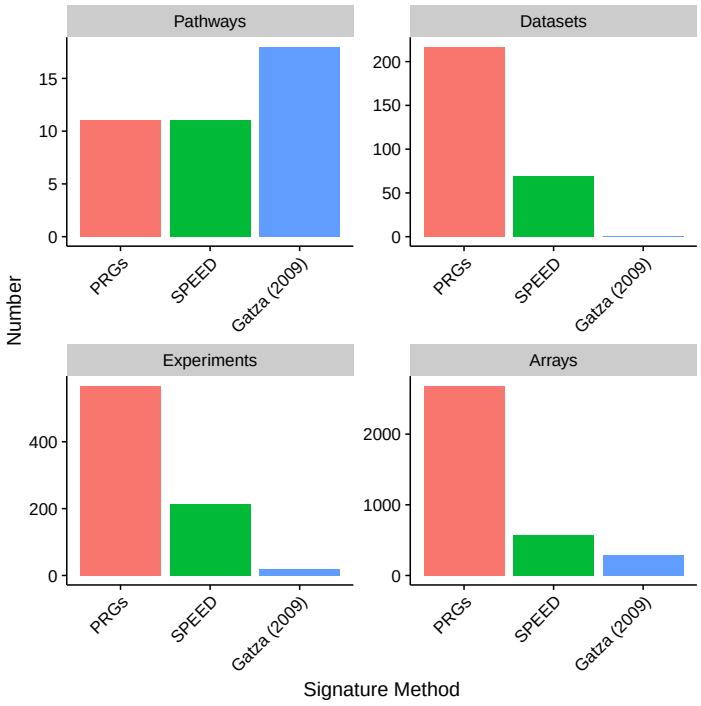

Figure 12: Comparison of dataset size between SPEED, Gatza (2009), and my Pathway-responsive genes (PRGs). (Gatza, Lucas, et al., 2010) derived 18 pathway signatures using only the MCF-10A cell line (thus also 18 experiments), and a total of 287 arrays. In 2014, they included additional signatures from other sources to a total of 53, but some are redundant, others not pathways, and all still limited to breast cancer (Gatza, Silva, et al., 2014). SPEED(Parikh et al., 2010) assembled consensus signatures for for 11 pathways using 69 GEO submissions, 215 different conditions and 572 arrays. My data set consists of 11 pathways, 217 GEO submissions, 568 different experiments and a total of 2687 arrays. This means I use more evidence per pathway and cover a broader set of experimental conditions, but also reflects the imposed limitation of only considering experiments if there are at least two unperturbed arrays available in order to estimate the basal variability. In addition, all my expression values are derived from raw data and not preprocessed data that I can not reproduce.

signatures is smaller but the number of overall experimental data much bigger (for a comparison, see figure 12).

I designed a pipeline to parse the experiment index, and download and process the corresponding experimental data as described in section 3.1.3. At the time, the release of the *ArrayExpress* R package was unable to process most of the experiments although the raw data as well as experiment description was available. In the spirit of increased reusability, I aimed at fixing the package instead of just writing my own processing routines, and the new version is available on BioConductor[2].

3.1.3 *From Experiments to Expression Data*

To go from an experiment accession identifier in the ArrayExpress database to the actual expression values of the arrays referenced in the curated data set, I planned to use the *ArrayExpress* package with its main function that should do exactly that: take an accession, and return an annotated `data.frame` with probe IDs in rows and samples in columns - with the annotation containing the metadata of the experiment. This was not working for the majority of experiments however, so I rewrote parts of the package and gave the patches to the original maintainers (for detailed list of changes, see the BioConductor VCS diff between 1.28.1 and 1.30.1). The update package has been downloaded by over 4,000 people since[3].

With the R package patched, I was able to download the associated data with all the Affymetrix accessions. I fitted a probe-level model (PLM) for each accession, calculating the `NUSE` and `RLE` quality control scores for each array. Arrays that had a `NUSE` or `RLE` over 0.1 I discarded as quality control failures. If less than two control and one perturbed array were remaining, I removed the whole experiment. If there were enough, I performed background correction, normalisation, and probe summarisation using the `rma` function in the *oligo* (Carvalho and Irizarry, 2010) or *affy* (Gautier et al., 2004) package (depending on whether the object was a `FeatureSet` or `AffyBatch`, respectively) using the corresponding platform design package. I then mapped the probe set identifiers to Human Genome Nomenclature Consortium (HGNC) symbols using the respective annotation packages (platform design and annotation package see table 4). I discarded probe identifiers mapping to more than one HGNC symbol, and averaged multiple probe sets that belonged to the same HGNC symbol using the `avereps` function in the *limma* package (G K Smyth, 2005). After processing and quality control, a total of 568 experiments and 2867 arrays remained, with a total of over 64 million single-gene measurements.

2 http://bioconductor.org/packages/release/bioc/html/ArrayExpress.html
3 http://bioconductor.org/packages/stats/bioc/ArrayExpress/

Table 4: Microarray platforms used in the curated experiments

ArrayExpress Platform ID	GEO Platform ID	Platform design package	Annotation package
A-AFFY-9	GPL91	pd.hg.u95a	hgu95a.db
A-AFFY-10	GPL92	pd.hg.u95b	hgu95b.db
A-AFFY-1	GPL8300	pd.hg.u95av2	hgu95av2.db
A-AFFY-33	GPL96	pd.hg.u133a	hgu133a.db
A-AFFY-34	GPL97	pd.hg.u133b	hgu133b.db
A-AFFY-44	GPL570	pd.hg.u133.plus.2	hgu133plus2.db
A-AFFY-37	GPL571	pd.hg.u133a.2	hgu133a2.db
A-GEOD-13667	GPL13667	pd.hg.u219	hgu219.db
A-AFFY-141	GPL6244	pd.hugene.1.0.st.v1	hugene10sttranscriptcluster.db
A-GEOD-11532	GPL11532	pd.hugene.1.1.st.v1	hugene11sttranscriptcluster.db
A-AFFY-141	GPL6244	pd.hugene.2.0.st	hugene20sttranscriptcluster.db
A-GEOD-17692	GPL17692	pd.hugene.2.1.st	hugene21sttranscriptcluster.db
A-AFFY-143	GPL20188	pd.huex.1.0.st.v1	huex10sttranscriptcluster.db
A-GEOD-16209	GPL16209	pd.huex.1.0.st.v2	huex10sttranscriptcluster.db
A-AFFY-76	GPL3921	pd.ht.hg.u133a	hthgu133a.db
A-GEOD-13158	GPL13158	pd.ht.hg.u133.plus.pm	?
A-GEOD-17586	GPL17586	pd.hta.2.0	hta20sttranscriptcluster.db

3.2 A MODEL FOR PERTURBATION-RESPONSE EXPERIMENTS

3.2.1 *Building a linear model from z-scores*

With the curated experiments, one requirement was that there were at least two arrays available that represent the control condition. This way, I can use the distribution of the control condition of each gene symbol as the reference distribution, and calculate the number of standard deviations that the average perturbed sample lies above or beneath its mean:

$$z_i = \frac{\overline{x}_i - \mu_i}{\sigma_i}$$

where i is the index of the gene, z_i is the obtained z-score, $\bar{x}_i$ is the mean perturbed expression level, μ_i is the mean expression level of the reference condition and σ_i its standard deviation. In addition, I use a `loess` model to smooth the relationship between a gene's expression level and its variance across samples as described in (Parikh et al., 2010). From the z-scores of all experiments and all pathways, I performed a multiple linear regression with the pathway as input and the z-scores as response variable for each gene separately:

$$Z_g \sim M \qquad \forall g \in genes$$

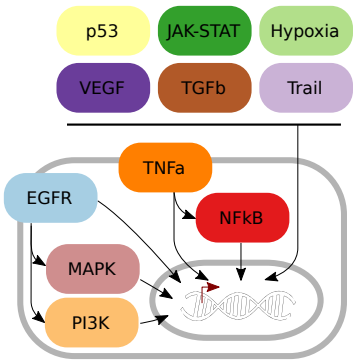

Figure 13: Structure of the perturbation-response model. For the multiple linear regression, I set the coefficients or perturbed pathways to 1 if a pathway was perturbed, 0 otherwise. In addition, EGFR perturbation also had MAPK and PI3K coefficients set, and TNFa had NFkB set.

Where Z_g is the z-score for a given gene g across all input experiments (as a column vector of experiments). M is a coefficients matrix (rows are experiments, columns pathways) that has the coefficient 1 if the the experiment had a pathway activated, -1 if inhibited, and 0 otherwise. For instance, the Hypoxia pathway had experiments with low oxygen conditions set as 1, HIF1A knockdown as -1, and all other experiments as 0. The same is true for EGFR and EGF treatment vs. EGFR inhibitors respectively, with the additional coefficients of MAPK and PI3K pathways set to 1 because of known pathway cross-talk. An overview of all modelled cross-talk is shown in figure 13.

$$
\begin{array}{cc}
 & \textit{pathway index} \\
\begin{matrix} z_g^{E_1} \\ z_g^{E_2} \\ z_g^{E_3} \\ z_g^{E_4} \\ z_g^{E_5} \\ \vdots \end{matrix} & \sim \begin{matrix} 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \\ 0 & 1 & 0 & \\ 0 & 1 & 1 & \\ 0 & 0 & 1 & \\ \vdots & & & \ddots \end{matrix}
\end{array}
\qquad \forall g \in \textit{genes}
$$

From the result of the linear model, I selected the top 100 genes per pathway according to their p-value and took their estimate (the fitted z-scores) as coefficient. I set all other gene coefficients to 0, so this yielded a matrix with HGNC symbols in rows and pathways in columns, where each pathway had 100 non-zero gene coefficients (cf. figure 11).

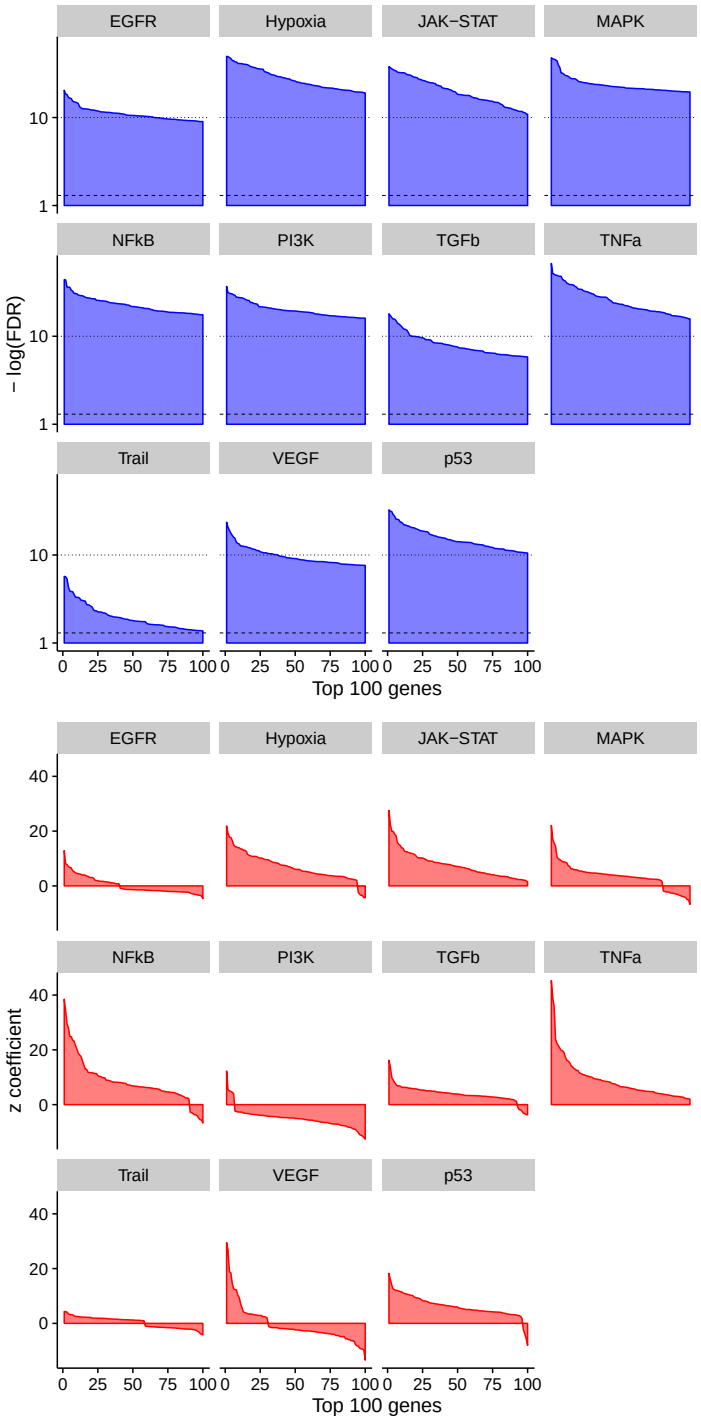

Figure 14: Distribution of the top 100 genes used in the model, as outcome for the multiple regression of all pathways. Top: Distribution of FDR-corrected p-values for signature genes (double log). Genes (horizonal axis) are ordered by significance. Dashed line at 5% FDR, dotted line at 10^{-10}. Bottom: Distribution of z-scores of the top 100 significant genes (different order compared to a). Signature genes are comprised of both up- and downregulated genes for most of the pathways. In no pathway a single or a few z-scores are high enough to overshadow the rest of the signature, indicating that the model is numerically stable.

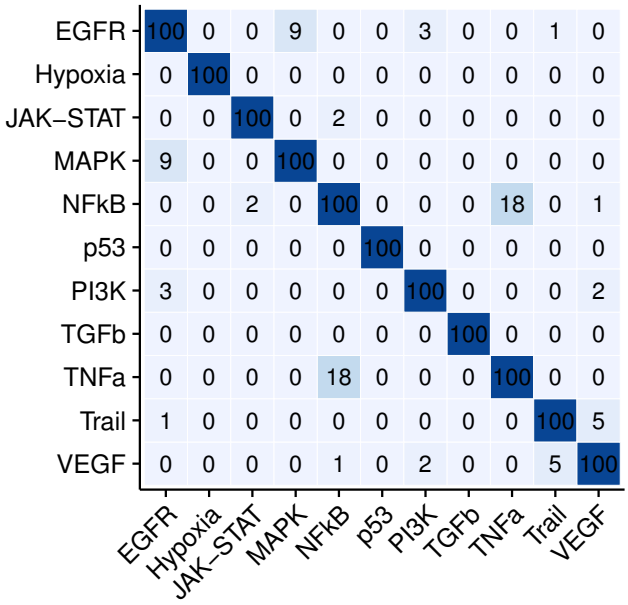

Figure 15: Overlap of signature genes for different pathways. For the 100 genes I selected for each individual pathway, this shows how many of those signature genes are present in another pathway as well. The overlap is generally minimal, with the highest numbers between TNFa and NFkB (18 of 100) and EGFR and MAPK (9 of 100). As both of these pathway pairs have one component that is directly upstream of the other, this is to be expected. It also shows that the response genes are specific to their perturbation and not a common phenotype like stress response that happens with any perturbation.

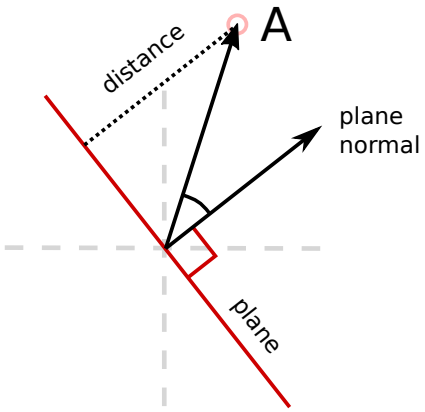

Figure 16: 2-dimensional schema of a distance calculation of a point from a plane by means of a dot product. The plane is defined by its normal vector, and the distance between it and the point marked A is calculated by using the formula $\vec{n} \cdot \vec{A}$.

3.2.2 *Computing pathway scores*

For each pathway signature I obtain an N-dimensional vector in gene expression space Z_i. This vector defines not only a direction, but also a plane on the origin of a coordinate system whose normal it represents. The distance of each sample (A) from that plane corresponds to the activation state of the pathway, calculated as the dot product between Z_i and the gene expression of a given sample (or fold change). A schema of this is shown in figure 16.

While the above only explains how to calculate the score for one pathway in one sample, I have 11 pathways and M samples. Each column in the matrix of perturbation-response genes (Z with N genes in rows and M samples in columns) corresponds to a plane in gene expression space (Z_i with N genes, $i \in M$). For samples, I have a gene expression matrix E with genes in columns (N dimensions) and samples in rows (M dimensions). I can now calculate all the dot products between Z_i and E_{sample} by performing a matrix multiplication to yield P, a matrix with all pathway scores (rows, indexed by i) in all samples (columns, M dimensions):

$$P = E * Z$$

This approach is vastly simpler and more computationally efficient than any other way of computing scores for other pathway methods that I compare in section 3.3.2. As I am more interested in the relative activation status between samples than the absolute values that are not particularly meaningful, I scaled each pathway or gene set score to have a mean of zero and standard deviation of one across all samples.

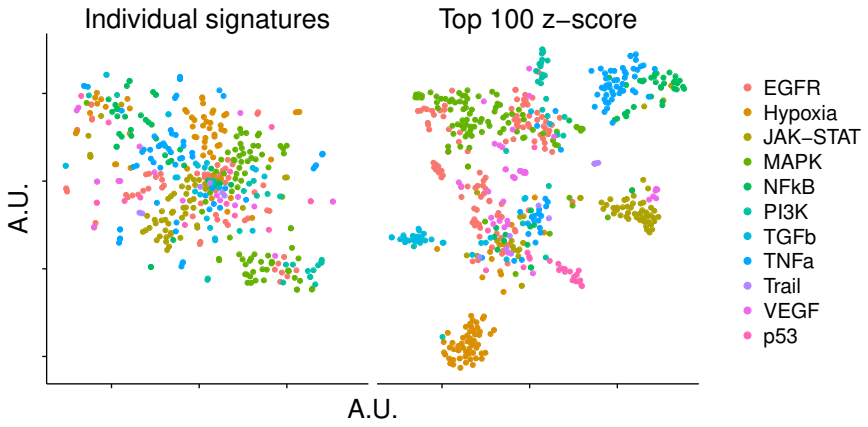

Figure 17: T-SNE plots for separation of perturbation experiments with different pathways in different colours. Fold changes of genes in individual perturbation experiments (10% FDR) do not cluster by pathway (left). Using a consensus signature of genes whose z-score is most consistently deregulated for each pathway instead, we can observe distinct clusters of perturbed pathways (right).

3.2.3 *Consensus gene signatures reveal pathway-response transcriptional modules*

Each experiment in my data set consisted of between 3 and 30 arrays, measuring between 8597 and 24,041 genes (72% over 15,000 genes). For extracting the individual signatures for each experiment, I calculated differential expression between the basal and perturbed condition and selected all fold changes that were significant at 10% FDR. If the pathway was inhibited, I took the negative fold changes. This yielded between 1 and 14,014 z-scores of significantly changing genes, reflecting the strength of the perturbation, experimental design and number of genes on the array. I assembled a matrix of significant fold changes by setting all non-significant or non-existent gene coefficients to 0, and use t-SNE (Van der Maaten and Hinton, 2008; Maaten, 2013) for dimensionality reduction on that matrix.

Looking at the fold changes in each perturbation experiment, I do not see an obvious clustering in which pathway was perturbed (figure 17, left). Distances between the fold changes of different experiments are about equal, with the slight exception of JAK-STAT and Hypoxia that are grouped to some extend, but the fold changes do not allow for a separation of pathways.

To compare this to my model, I calculated pathway scores from the unfiltered fold changes of my input experiments, or negative fold changes in the case of inhibitions. Applying the model building as described in section 3.2.1, I obtain an expression matrix with genes in rows and pathways in columns. As this is a linear model of the z-scores

of those different pathways, I can obtain activation scores (my inferred pathway activity) by a matrix multiplication between the model matrix and gene expression values from samples, as long as the identifiers are in the same order. Using only the top 100 significant genes (figure 14), I can show a separation of all pathways in groups (and some weak perturbations overlapping in the centre). However, these clusters only tell us that across all scores, individual pathways produce patterns distinct from others. It does not tell us the obtained scores actually correspond to the perturbed pathway.

I calculate linear associations using the `lm` function in R between a pathway coefficients matrix (experiments in rows and pathways in columns, with the value 1 where a pathway was activated, -1 where a pathway was inhibited, and 0 otherwise; here, no cross-talk was modelled explicitly) and the pathway scores obtained by using the different methods as described in sections 3.2.2 and 3.3.2. I then plot the result of the associations as a matrix (figure 18, left) where the perturbed pathway and the obtained pathway score are on the axes, and the colour (blue for positive and red for negative correlation) indicates the Wald statistic between them.

With the linear model I use for this purpose (implementation details in section 3.2.3), I find that all pathway scores are strongly associated with the perturbation they were derived from (figure 18, left), and some for a perturbation with a known pathway cross-talk (e.g. EGFR activating MAPK or TNFa activating NFkB).

As a method for inspection of the pathway scores of the individual experiments, I plot a heatmap using the R package *pheatmap* with the pathways scores in rows, experiments in columns, and the colour indicates the relative (column-scaled) activation of each pathway (figure 18, right). In order to make activations and inhibitions more comparable, I take the negative pathway score for each inhibition and annotate the columns separately indicating which pathway was perturbed, and which kind of perturbation it was. This way, I can show clusters of similar pathway activation patterns that should correspond to the pathways that were perturbed.

The above test showed us that the mean of assigned scores is significantly altered when a given pathway is perturbed, but it gives us only limited information about the heterogeneity in the individual experiments. To this end, I calculated the relative pathway activations for each individual experiment (details section 3.2.3) and show the results in a heatmap ordered by perturbation (figure 18, right). For easier interpretation, I show all scores in the direction of activation by taking the negative scores of inhibitions (so red indicates a match, not necessarily an activation) and indicating whether the original experiment was an activation or inhibition separately. We can clearly see that the

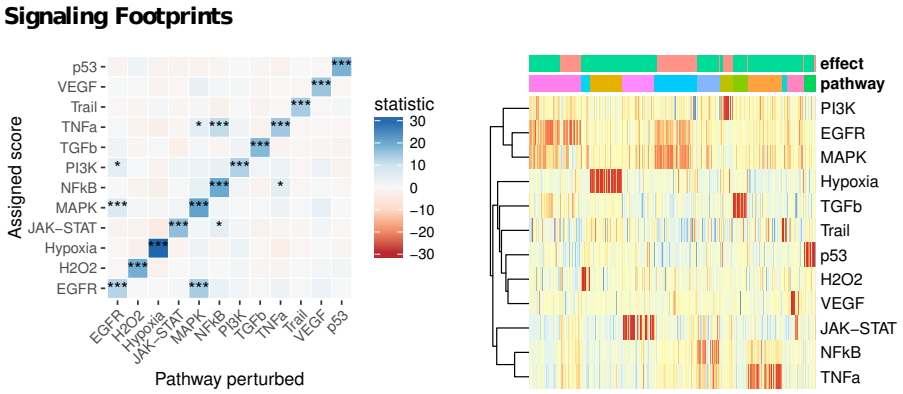

Figure 18: Recovery of the perturbations across pathways (left) and for individual experiments (right) using the consensus signature. Left: Associations calculated between perturbed pathways and the scores obtained by the model of pathway-response genes (PRGs). Along the diagonal each pathway is strongly ($p < 10^{-10}$) associated with its own perturbation. Significant off-diagonal elements are sparse and only show a strong association between EGFR/MAPK and TNFa/NFkB and a weaker association ($p < 10^{-5}$) between EGFR/PI3K and MAPK/TNFa. Right: Heatmap of relative pathway scores in each perturbation experiment. 523 experiments in columns, annotated with the perturbation effect (green for activation, orange for inhibition) and pathway perturbed (same order as b). Pathway scores in rows cluster between EGFR/MAPK and to a lesser extent PI3K, and TNFa/NFkB. Colour indicates activation or inhibition strength. Order of pathways: EGFR, H2O2, Hypoxia, JAK-STAT, MAPK, NFkB, PI3K, TGFb, TNFa, Trail, VEGF, and p53.

majority of pathway scores across experiments agrees with the pathway that was actually perturbed in the experiment.

Taken together, those two approaches imply that the perturbation-response signature is both correct on average, as well as in the majority of individual cases. For the remaining cases, I do not know if this is something my (admittedly very simple) model misses or the pathway was not sufficiently perturbed with the given parameters the experimentalists used (e.g. a low concentration of the perturbing agent, too little time for genes to change their expression significantly, or a general experimental error).

3.3 COMPARISON TO PATHWAY MAPPING METHODS

3.3.1 *Transcriptional footprints are fundamentally different to pathway expression*

As already indicated in my initial motivation (figure 10), commonly used pathway methods map the mRNA expression level of signalling molecules on the corresponding pathways to infer a pathway-level score.

But are those really different? Maybe the genes expressed upon pathway perturbation are feedback regulators of the same pathway. In this case, pathway mapping could very well identify perturbations because those are already incorporated in the pathway structure.

However, I find that the overlap of my pathway-responsive genes and known Gene Ontology categories or Reactome pathways is either small or non-existent, as shown in figure 19. Interestingly, I find that for the only pathway signature that is strongly tied to a specific inhibitor (MAPK and MEK inhibitors), the one gene to overlap between all methods is MEK itself. This effect of an inhibitor on the expression of its target gene has been shown before (Iskar et al., 2010).

Notwithstanding these differences, computing which unfiltered GO gene sets are over-represented in our signature genes may still give us additional insight into which lower-level processes (and not signalling pathways) are activated upon pathway stimulation. For this, I computed enrichment of Gene Ontology categories using a Fisher's exact test and all gene sets that had between 5 and 500 members.

What I find is largely expected, but it also goes to show that modulation of signalling pathways mainly triggers expression changes in biological processes related to the pathway and not the pathway members itself.

For instance (figure 20), EGFR and MAPK drive, while p53-mediated DNA damage response abrogates the cell cycle and DNA replication (especially the G1/S checkpoint (Di Leonardo et al., 1994)). In addition, MAPK and p53 have more specific opposing roles in components

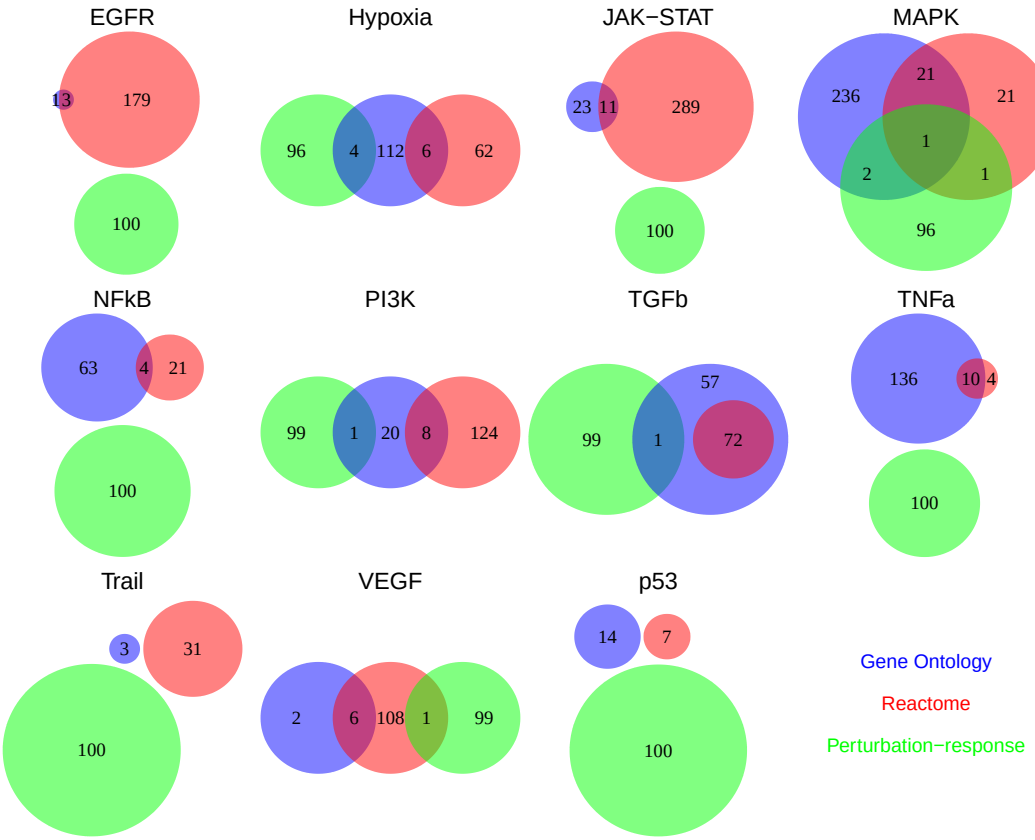

Figure 19: Difference in gene sets between Perturbation-response genes and Gene Ontology/Reactome. Perturbation-response genes are different to path- way members and gene annotations, with between 0 and 4 genes in common. However, pathway annotation is quite different from gene annotations here as well.

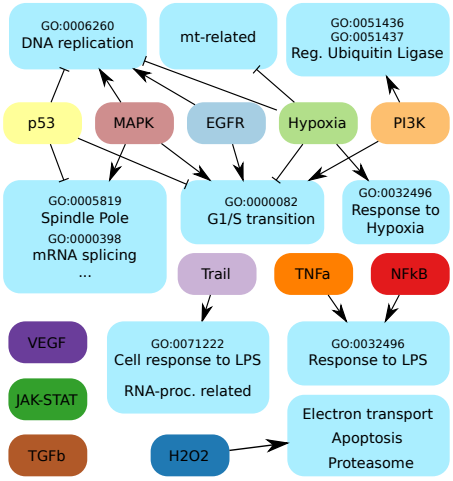

Figure 20: Functional annotations of perturbation-response genes

Table 5: Mapping of pathways to Gene Ontology Categories

2*Pathway	Gene Ontology	
	ID	Name
EGFR	GO:0007259	ERBB signaling pathway
Hypoxia	GO:0071456	cellular response to hypoxia
JAK-STAT	GO:0007259	JAK-STAT signal transduction
MAPK	GO:0000165	MAPK cascade
NFkB	GO:0038061	NIK/NF-kappaB signaling
p53	GO:0030330	DNA damage response, signal transduction by p53 class mediator
PI3K	GO:0014065	phosphatidylinositol 3-kinase signaling
TNFa	GO:0033209	tumor necrosis factor-mediated signaling pathway
TGFb	GO:0007179	transforming growth factor beta receptor signaling pathway
Trail	GO:0036462	TRAIL-activated apoptotic signaling pathway
VEGF	GO:0038084	vascular endothelial growth factor signaling pathway

of the spindle pole or mRNA splicing (among others). PI3K stimulation also promotes G1/S transition, as well as regulation of ubiquitin. Hypoxia on the other hand inhibits the expression of G1/S genes and mitochondria-related genes (where oxidative phosphorylation happens (Schultz and Chan, 2001)). TNFa/NFkB show enrichment in the category of genes responsive to known pathway activator LPS (P A Baeuerle and Henkel, 1994), and so does Trail (Halaas et al., 2000).

3.3.2 *Computing pathway scores for other methods*

Pathway and Gene Ontology scores

I matched my defined set of pathways to the publicly available pathway databases Reactome (Croft et al., 2011) and KEGG (Kanehisa and Goto, 2000; Minoru Kanehisa et al., 2009), and Gene Ontology (GO) (Ashburner et al., 2000; Gene Ontology Consortium, 2004) categories. In order to make them comparable, I selected the same set of pathways across different resources. I calculated pathway scores for these sets using GSVA as described in section 2.1.2.

The pathways I used are shown in table 5 for Gene Ontology, table 7 for Reactome (used in Reactome enrichment and Pathifier), and KEGG in table 6 (used in Signaling Pathway Impact Analysis).

SPIA scores

Signaling Pathway Impact Analysis (SPIA) (Tarca et al., 2008) is a method that utilizes the directionality and signs in a KEGG pathway graph to determine if in a given pathway structure the available species are more or less able to transduce a signal. As the species considered for a pathway are usually mRNAs of genes, this method still infers

Table 6: Mapping of pathways to SPIA (KEGG pathways)

2*Pathway	SPIA	
	ID	Name
EGFR	04012	ErbB signaling pathway
Hypoxia	-	-
JAK-STAT	04630	Jak-STAT signaling pathway
MAPK	04010	MAPK signaling pathway
NFkB	04064	NF-kappa B signaling pathway
p53	-	-
PI3K	04150	mTOR signaling pathway
TNFa	-	-
TGFb	04350	TGF-beta signaling pathway
Trail	04210	Apoptosis
VEGF	04370	VEGF signaling pathway

signalling activity by proxy of gene expression. In order to do this, SPIA scores require the comparison with a normal condition in order to compute both their scores and their significance.

I used the *SPIA* BioConductor package (Tarca et al., nodate) in my analyses that implements an updated version of the 2008 algorithm. I calculated my pathway scores either for each cell line compared to the rest of a given tissue (for the GDSC and drug response data) or compared to the tissue-matched normals (for the TCGA data used in driver and survival associations).

Pathifier scores

As with SPIA, Pathifier (Drier, Sheffer, and Domany, 2013) requires the comparison with a normal condition in order to compute its scores. The difference is that it infers pathway structure from gene expression data itself instead of relying on prior knowledge.

I used the *Pathifier* BioConductor package (Drier, Sheffer, and Domany, 2013) in my analyses. I calculated pathway scores either for each cell line compared to the rest of a given tissue (for the GDSC and drug response data) or compared to the tissue-matched normals (for the TCGA data used in driver and survival associations).

PARADIGM scores

I downloaded the PARADIGM software from the public software repository[4] and a model of the cell signalling network published in (The Cancer Genome Atlas Research Network et al., 2013) from the corresponding TCGA publication[5]. I used my `voom`-transformed RNA-seq gene expression data that I normalized using ranks to assign equally

4 https://github.com/sbenz/Paradigm
5 https://tcga-data.nci.nih.gov/docs/publications/coadread_2012/

Table 7: Mapping of pathways to Pathifier (Reactome pathways)

1*Pathway	Name
2*EGFR	Signaling by EGFR
	Signaling by EGFR in Cancer
Hypoxia	Cellular response to hypoxia
3*JAK-STAT	Signaling by Interleukins
	Interferon Signaling
	Signalling to STAT3
MAPK	Signalling to ERKs
2*NFkB	TAK1 activates NFkB:w: by phosphorylation and activation of IKKs complex
	RIP-mediated NFkB activation via ZBP1
p53	Transcriptional Regulation by TP53
4*PI3K	PI3K Cascade
	Constitutive Signaling by Aberrant PI3K in Cancer
	PI3K/AKT Signaling in Cancer
	PI3K/AKT activation
TNFa	TNF signaling
TGFb	Signaling by TGF-beta Receptor Complex
Trail	TRAIL signaling
VEGF	Signaling by VEGF

spaced values between 0 and 1 for each sample within a given tissue, as recommended in the manual. I then ran PARADIGM inference using the same options as in the above publication for each sample separately. I used nodes in the network representing pathway activity to our set of pathways to obtain pathway scores that are comparable to the other methods, averaging scores where there were more than one for a given sample and node.

3.3.3 *Comparison within perturbation experiments*

My first investigation is to which extent pathway expression is correlated with known perturbations. They would be if the perturbation causes feedback cycles where stimulation (or inhibition) of a pathway will lead to changes in the expression of some of its own components, such as inhibitors in the case of negative feedback.

But is this enough of a change to determine pathway activity from the gene expression of its components? If so, it would enable the expression to be used as an actionable biomarker for modulating the pathway's activity, e.g. by means of treatment using a kinase inhibitor. If not, we can still use it as a measure of pathway expression, but not make statements about its activity or whether adding a drug would change its state.

For methods based on Gene Set Enrichment Analysis (Reactome, BioCarta, Gene Ontology), the associations they provide between a

Table 8: Mapping of pathways to PARADIGM nodes

2*Pathway	2*PARADIGM
EGFR	epidermal growth factor receptor activity (abstract)
Hypoxia	response to hypoxia (abstract)
JAK-STAT	STAT-1-3-5-active
1*MAPK	MEK-1-2-active
NFkB	NFkB Complex (complex)
p53	response to DNA damage stimulus (abstract)
PI3K	PIK3CA
TNFa	tumor necrosis factor receptor activity (abstract)
TGFb	SMAD1-5-8-active
Trail	induction of apoptosis (abstract)
VEGF	platelet-derived growth factor receptor activity (abstract)

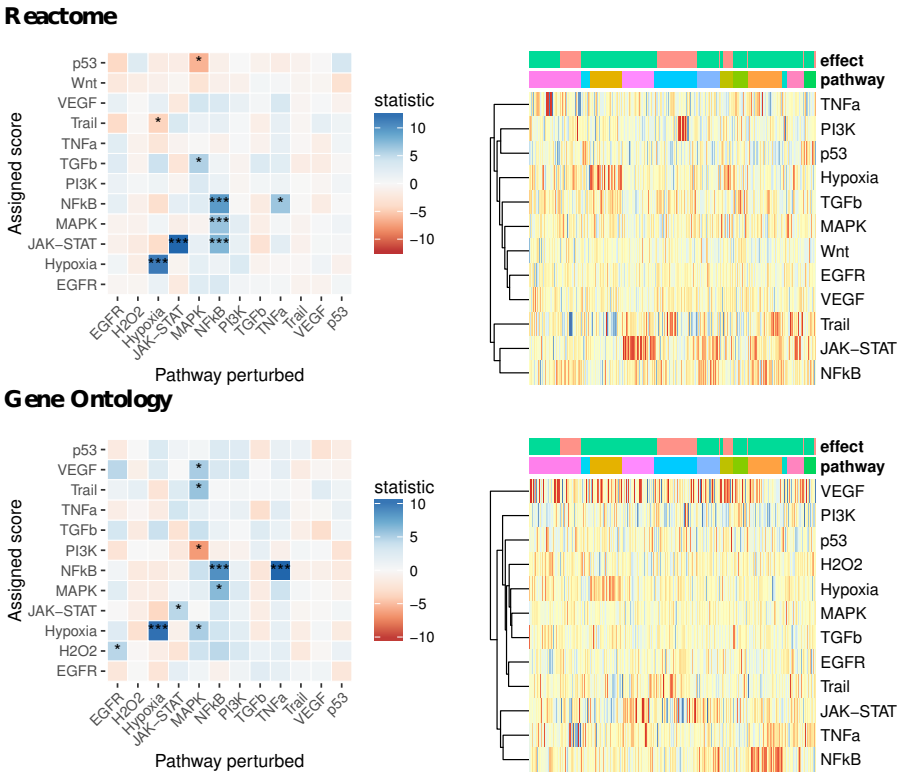

Figure 21: Comparison to pathway member GSVA. Significant associations (marked with $*$ for $p < 10^{-5}$ and *** for $p < 10^{-10}$) between a pathway perturbation and a higher (blue) or lower (red) pathway score are shown on the left, pathway scores for individual experiments on the right (order of pathways: EGFR, H2O2, Hypoxia, JAK-STAT, MAPK, NFkB, PI3K, TGFb, TNFa, Trail, VEGF, p53)

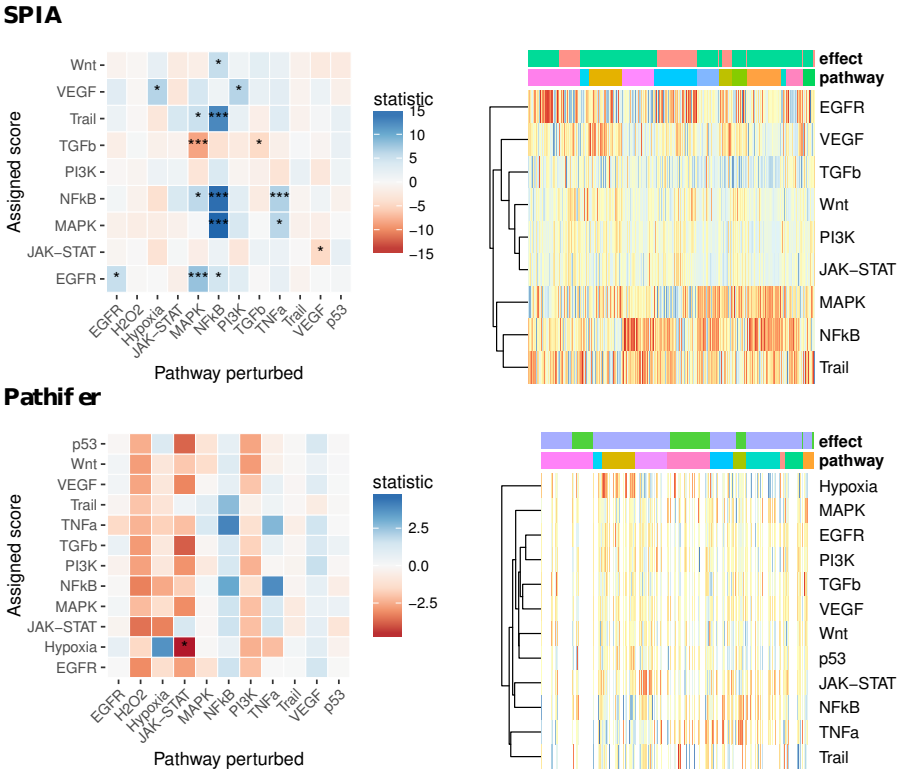

Figure 22: Comparison to other pathway methods, legend as in figure 21

perturbed pathway and a significant change in score (figure 21) are very much different from the pathway scores obtained from the consensus of the response itself (figure 18). Reactome shows the strongest associations for the perturbation and pathway score for Hypoxia and JAK-STAT (followed by NFkB, MAPK, and JAK-STAT pathway expression upon NFkB perturbation). Gene Ontology shows strong self-associations for Hypoxia and NFkB, and NFkB expression changes upon TNFa stimulation.

These results provide a further confirmation that pathway expressed as defined as a consensus score of pathway members is fundamentally different to pathway activity and can not be used as a measure for such. All the categories that provide strong associations are explicitly curated to contain the pathway response, as opposed to the members. This is true for a "Response to Hypoxia" for Reactome and Gene Ontology (cf. tables 7 and 5), while the BioCarta pathway that does not provide an association focussed more on the pathway itself. A similar pattern can be observed for TNFa/NFkB and the "Response to LPS" category (also tables 7 and 5).

SPIA (figure 22, top) seems to assign more extreme relative scores to a couple of pathways (MAPK, EGFR, NFkB, Trail) no matter which pathway was perturbed. It does not show a clear unique link between a pathway stimulation and its impact on scores in the same pathway.

Table 9: Area under the ROC curve (AUC) for methods and pathways in figure 23. Best performance in bold.

2*Pathway	Method				
	Perturbation-response	Gene Ontology	Reactome	SPIA	Pathifier
EGFR	**0.84**	0.38	0.50	0.64	0.50
Hypoxia	**0.95**	0.86	0.86	-	0.71
JAK-STAT	**0.80**	0.76	0.50	0.43	0.76
MAPK	**0.83**	0.59	0.46	0.29	0.48
NFkB	0.77	0.56	0.75	**0.79**	0.60
PI3K	**0.81**	0.26	0.47	0.35	0.46
TGFb	**0.94**	0.68	0.75	0.15	-
TNFa	**0.87**	0.70	0.53	-	0.63
Trail	**0.91**	0.42	0.56	0.52	0.39
VEGF	**0.76**	0.58	0.52	0.52	0.52
p53	**0.86**	0.32	0.67	-	0.48

Pathifier (figure 21, bottom) fails to derive scores for the majority of perturbation experiments (their method tries to construct principal curves from control-experiments that need at least three arrays in that condition) and for the remaining does not provide any strong associations (with the strongest being the up-regulation of the hypoxic response upon JAK-STAT stimulation, otherwise all $p > 10^{-5}$).

3.3.4 *Comparison across perturbation experiments*

The previous section focussed on pathway activation in relative terms (whether, given a pathway perturbation, the score assigned to any one pathway is higher or lower compared to other pathways). I am now interested in how well pathway methods are able to assign activity scores across different experiments. I use as input data the fold changes (difference in *log* space, or negative difference if inhibition) of the curated perturbation experiments, and see how well the pathway scores derived from those are ordered for each pathway across experiments. This is meant as a comparison of how well the pathway scores obtained by different methods correspond to perturbations, and not to assign significance of the perturbation-response signature model.[6]

The standard non-parametric (*i.e.*, only based on order rather than value) of quantifying this is the area under Receiver-Operator (ROC) curves, the results of which is shown in figure 23 (AUCs in table 9). For all pathways except NFkB and JAK-STAT (where multiple methods are tied), the consensus gene signature I derived is also bet-

6 Because then I would need an independent set of experiments. This is not required as I assessed the significance of the genes in the model analytically, with results shown in figure 14. The comparison here serves to show that my model corresponds better to perturbations.

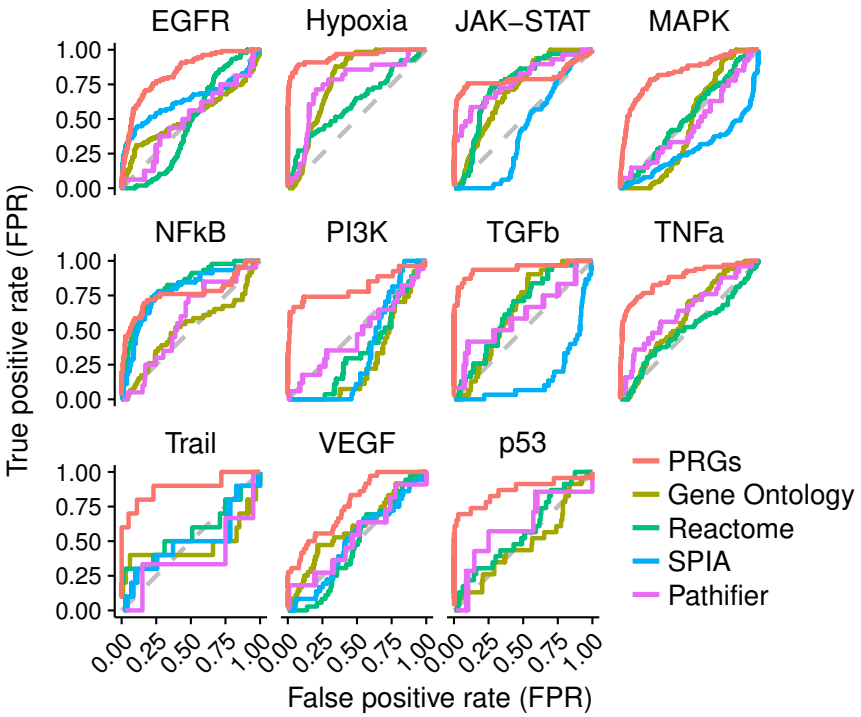

Figure 23: Perturbation recall on input experiments as a comparison across methods for each pathway. ROC curves for different methods ranking perturbation experiments by their pathway score. Pathway-response genes (PRGs) show better performance for all pathways except JAK-STAT and NFkB, where other methods are equal. Gene Ontology and Reactome scores are obtained by Gene Set Variation Analysis (GSVA). Pathifier using Reactome gene sets. Performance is indicated with the area under each curve, values indicated in table 9. Lines for methods are only drawn if there were at least five experiments where a given pathway was perturbed and the method could derive scores.

ter able to detect pathway activations across experiments. VEGF is not well recovered by any method, potentially due to the relatedness with EGFR/MAPK and the low number of experiments. Interestingly, SPIA in 4 out of 11 cases systematically assigns pathway scores in the wrong direction of the experimental perturbations (this may be due to increased expression of negative feedback regulators that SPIA picks up due to the underlying directed KEGG graph).

3.4 DISCUSSION

3.4.1 *Need for an improved database and model*

As I have outlined already in the discussion of chapter 2, the original SPEED implementation suffered from a multitude of arbitrary cutoffs (z-score, overall expression, highest expressed fraction of genes, whether the signature genes are allowed to overlap between pathways), highly correlated gene sets (that could be fixed by optimising those cutoffs), discarding most of its information (not taking into account down-regulated genes in general, modestly but consistently upregulated genes, etc.), and potentially weak input experiments (mostly processed data without any knowledge of how this was done, calculating z-scores for only one control condition, and never thoroughly assessing the quality of the model).

Also the results that I obtained in terms of drug associations were significant but not stellar. This, combined with the fact that the original database was curated in 2010 where there were fewer perturbation experiments available, is an indication that building a new database could result in a much improved model.

3.4.2 *Value of curated perturbation experiments*

The average entry in the ArrayExpress database comes with a fair level of curation. Newer entries are often linked to ontology terms in the Experimental Factor Ontology (Malone et al., 2010) or others. Linking those to pathways, however, is non-trivial because the available information is not accessible via an API. This is only possible using the Gene Expression Atlas (Parkinson et al., 2007; Helen Parkinson et al., 2009), but it also requires a much higher level of curation than the experiments that are just deposited in the database and hence the number of experiments that made it into the Atlas is much smaller. In order to interrogate cellular systems (as those provided by the GDSC and the TCGA in this study), I needed a high number of perturbation experiments that are not available in the Atlas yet. This required a focussed curation of those pathways.

The value of this approach, combined with putting it into a machine readable format that includes metadata (as opposed to the original SPEED database that came either as a collection of un-annotated text files or SQLite database), allows for easy reuse in other projects and extension by adding additional experiments or pathways.

3.4.3 *Importance of distinguishing between pathway expression and activity*

Seeing how my model compares to other pathway methods that map mRNA expression to signalling molecules in one way or another underlines the importance of distinguishing between pathway expression and activity and, by extension, that the expression of pathway members should not be used to make inferences about activity. This is irrespective of using the pathway as a simple gene set or a directed graph. Pathway expression alone does not tell us much about the involved biology (a more expressed pathway likely only has an impact on the cell if it is more active as well.

However, this is a jump in reasoning that we can not make for the reasons outlined above. Also, pathway expression is not actionable by itself, because there is no straightforward way to change the expression of a pathway instead of its activity (as does a drug, for instance).

It can of course be argued that there are feedback loops where the stimulation of a pathway triggers changes in expression in its components (as I find with JAK-STAT and Hypoxia), either in the same (positive feedback) or the opposite (negative feedback) direction.

This way of thinking makes sense only for positive feedback on the same pathway: if it is negative feedback we would see an incorrect decrease of expression or a increase of associated inhibitors (I see this kind of correlation using SPIA and how it incorrectly orders experiments in figure 23), or if it is on another pathway we would assign significance to the expression change of this pathway although this is merely a downstream consequence that did not cause the phenotype we observe at all.

I can not claim that my model will always find causal aberrations. But I do argue that it is more likely for it to find them compared to other pathway methods, because the way it infers pathway activity from gene expression is closer to the actual activity by definition. If this is true, I would expect it not only to be able to recover known perturbation experiments better (figures 21 and 22) but also provide more insight into which pathways driver activate mutations and influence drug sensitivity due to oncogene addiction. Now that I have a model that I know corresponds to pathway perturbations in a wide range of conditions I can actually put those statements to the test.

4

FUNCTIONAL EVALUATION OF PATHWAY METHODS IN BASAL GENE EXPRESSION

This chapter represents the evaluation of the model proposed in the previous one. It aims to investigate the functional relevance of pathway-responsive genes (PRGs) compared to other pathway methods. In terms of outcomes, I am interested in how well the available platform is able to explain:

- The signalling activity mediated by driver mutations and copy number aberrations
- How this affects sensitivity to targeted therapies
- Which pathways have implications for patient survival

The chapter represents the results stage of improved pathway signatures. All analyses, plots, and written text in this thesis I produced myself, while incorporating comments from the coauthors:

Schubert M, Klinger B, Klünemann M, Garnett MJ, Blüthgen N, Saez-Rodriguez J. "*Perturbation-response genes reveal signaling footprints in cancer gene expression.*" **bioRxiv** 065672 (2016). doi:10.1101/065672

4.1 PERTURBATION-RESPONSE SIGNATURES FOR BASAL GENE EXPRESSION

4.1.1 *Correlation in basal expression*

For all pathway scores described in sections 3.2.2 and 3.3.2 using basal gene expression, I calculated the Pearson correlation across all samples and tissues, and plotted the correlation matrix between each pathway combination using the R package *corplot* for GDSC and TCGA data separately. The correlation between a pathway and itself is 1 by definition, all other values range from -1 (where the values are identical except for a constant factor $x_1 = k * x_2 + \varepsilon$ where $k < 0$ and residual $\varepsilon = 0$) to 1 (same for $k > 0$).

4.1.2 *A linear model is easily transferable*

As I am most interested in how basal signalling activity in cancer cell lines and tumour samples correlates with outcomes such as drug response and survival, one crucial point of deriving pathway response

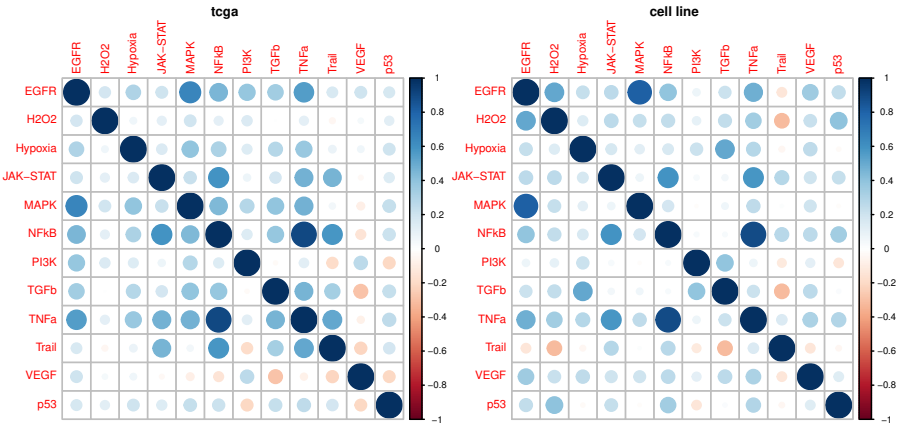

Figure 24: Correlation for TCGA primary tumour data (left) and GDSC cell lines (right) for perturbation-response genes in basal expression. Positive correlation in blue, negative in red. Circle size and shade correspond to correlation strength. Pathways that showed cross-activation with pathway perturbations are more highly correlated in basal expression as well.

signatures is that they also need to reflect pathway activation in basal gene expression. At this point, I know that my perturbation-response signature genes correspond to the pathway activation between a basal and perturbed condition, but not how well the pattern of pathway activation from those perturbation experiments corresponds to gene expression footprints of a constitutively active signalling pathway leaves in basal gene expression.

Providing a definitive proof that the same signatures obtained correspond to constitutive activation in basal expression is only possible using a range of functional readouts, *e.g.* mutations activating pathways and drug response on oncogene addiction.

What I can more easily show, however, is how well known pathway cross-talk and structure (as is known in literature and I have shown in figures 13 and 18) translates to pathway correlation in basal expression, and how well they agree between tumours and cell lines. To this end, I computed pathway scores for all primary tumours in the TCGA and cell lines in the GDSC (details section 4.1.1), and checked how well they correlate.

I find that not only are the correlations in basal gene expression (figure 24) similar to the ones in perturbed experiments (figure 18), but they also correlate very well with known cross-talk and do not change more than could be biologically expected between primary tumours and cell lines. I find a high correlation between EGFR and MAPK and a lesser one to PI3K, as well as TNFa and NFkB. Another high correlation is between NFkB and JAK-STAT that I also observed in perturbation experiments (below cutoff of 10^{-5} of figure 18).

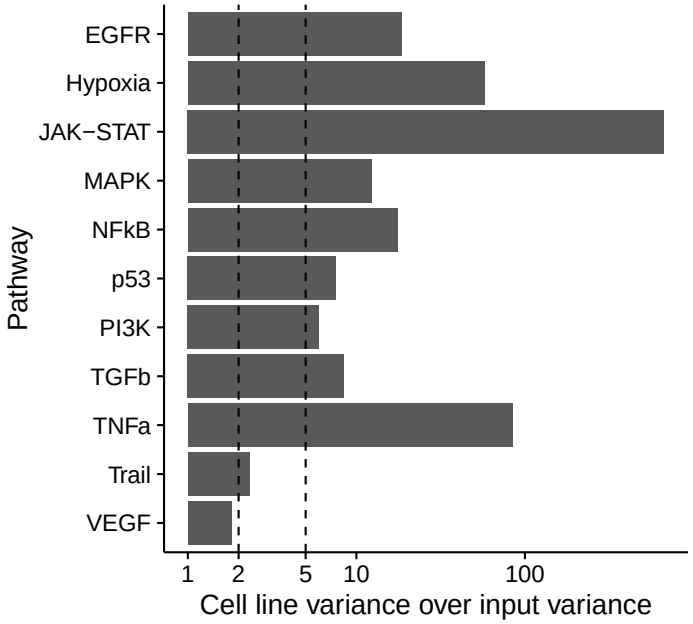

Figure 25: Stability of basal pathway scores when bootstrapping input experiments. The variance of pathway scores in cell lines given bootstraps is more than five times as high compared to the variance of bootstraps given cell lines for all pathways except two (Trail and VEGF), where it is roughly twice as high.

4.1.3 *Dependence on input experiments*

At this point I found strong indications that the gene expression changes in perturbation experiments can be described across many conditions using a linear consensus signature whose score reflects pathway activation and this signature is translatable to basal pathway activity in non-perturbed gene expression.

What I have not shown yet is how stable the whole process is, *i.e.* much the selection of input experiments influences the scores in basal expression. In order to investigate this, I bootstrapped the experiment selection 1,000 times and built models as described in section 3.2.1, using the GDSC basal expression and yielding a 3-dimensional scores tensor with a score for each cell line, pathway, and bootstrap. I regressed out either the effect of the individual cell lines or the effect of the bootstraps, and quantified the relative remaining variance in the scores:

$$ratio = \frac{var(cell\,lines \mid bootstraps)}{var(bootstraps \mid cell\,lines)}$$

The overview of those relative variances is shown in figure 25. Here, a value smaller than 1 indicates that the experiment selection has a bigger effect than the cell line I apply the model to (*i.e.*, we can not observe a

biological effect because it is hidden by the noise of input experiments) and values bigger than 1 indicate the times the biological context is more influential for providing the score than the input selection (i.e, the biological context outweighs the exact selection of perturbation input). For most pathways (TNFa, NFkB, MAPK, JAK-STAT, Hypoxia, and EGFR) the variance of the input experiments was only a negligible influence (smaller than 10%) on the final scores. For p53, TGFb, and PI3K selection of the input experiments still accounts for under 20% of the observed variance of basal pathway scores. For Trail and VEGF, it is about twice as large as the variance of the input experiments. This can either point to the pathway not being particularly well defined (if the experiments are also not well recovered by the signature, like for VEGF in figure 23), or that the overall number of input experiments is too small to guarantee a stable signature (like for Trail, where the signature performs well but the overall number of experiments are low, cf. figure 23).

4.1.4 *Pathway scores for other methods*

I used the Firehose tool[1] (release `2016_01_28`) from the BROAD institute to download data labelled level 3 RNA-seq version 2 (files names including *RSEM_genes__data.Level_3*, unpacked them, selected all files contained in the archive that contained the name *rnaseqv2*) for all cancer types for which it was available. I extracted the raw counts from each of the text files for each gene. I then performed a `voom` transformation (R package *limma* (G K Smyth, 2005)) for each TCGA study separately. The result of this transformation are expression values that I can use linear modelling techniques on, unlike the raw RNA-seq read counts. With the resulting expression matrix, I selected only genes that had an HGNC symbol associated with them. I used the function `avereps` (R package *limma*) to average rows that had the same gene symbol. This resulted in one expression matrix per TCGA tissue, for a total of 34 tissues including 9179 tumour and 661 matched normal samples.

I computed the pathway scores for perturbation-response genes as well as all other method using the subset of tissues in the TCGA gene expression data that had 30 or more normals in the same tissue (TCGA identifier `11A`) to compare. This is to make full use of the advanced pathway methods (figure 27). For the cell lines in the GDSC, I do not have tissue-matched normals, so this is not possible and I calculate the scores for the cell lines alone or comparing one cell line to the rest of the tissue if the method requires it (SPIA and Pathifier).

1 http://gdac.broadinstitute.org/

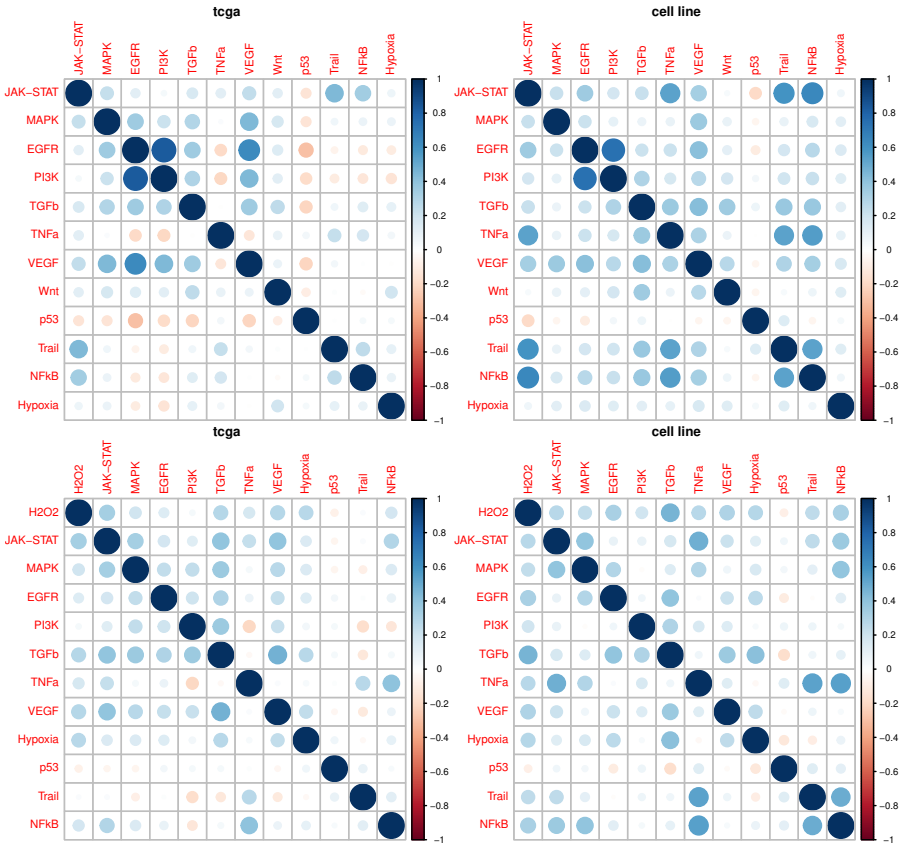

Figure 26: Comparison to pathway member GSEA for TCGA primary tumour data (left) and GDSC cell lines (right) for Reactome (top) and Gene Ontology (bottom). Sign of the correlation indicated by colour of the points (positive correlation blue, negative correlation red), strength by colour shade and size of the point.

For the gene set methods (figure 26), the amount of correlation between pathway scores they provide is in the same order of the one for pathway-response genes. Reactome (top) shows increased correlation between EGFR and MAPK (to a lesser extent PI3K), and between TNFa and NFkB (and to a lesser extent also JAK-STAT and Trail). Gene Ontology (bottom) is very similar to Reactome. For both methods, the correlation between pathways is similar from primary data (TCGA, left panel of figures) to cell lines (GDSC, right panel).

In comparison to the the pathway-responsive genes (figure 24) and the gene set methods (26), SPIA and Pathifier had more data available to compute their scores for the TCGA cohorts because they made use of the tissue-matched normals. This has a minor impact on the correlation between pathway scores obtained by SPIA (both show a high correlation between MAPK, NFkB, and Trail, top panel), but a major impact on the ones obtained by Pathifier (middle): here, supplying normals cause all the pathway scores to be highly correlated, while there is almost no correlation to be observed for the cell lines. For PARADIGM (bottom), there is almost no correlation between the nodes in the inference graph that correspond to pathway activity.

4.2 RECALL OF KNOWN PATHWAY MODIFIERS

4.2.1 *Pathway scores and mutations/CNAs*

I computed pathway scores for TCGA cohorts where there were tissue-matched normals available. For mutated genes, I considered all genes that had a change of coding sequence (SNP, small indels in MAF files) as mutated and all others as not mutated. For copy number alterations, I used the thresholded GISTIC scores, where we considered homozygous deletions (-2) and strong amplifications (2) as altered, no change (0) as basal and discarded intermediate values (-1, 1) from my analysis. I focussed on the subset of 464 driver genes that were also used in the GDSC. I used the sets of mutations and CNAs to compute the linear associations between samples for all different methods I looked at. I did not regress out the cancer type in order to keep associations where mutations/CNAs are highly correlated with it, but highlighted all associations that passed the significance threshold of $FDR < 0.05$ (for each pathway method individually) after such a correction.

4.2.2 *Associations using driver mutations and CNAs*

If my reasoning is correct and pathway-response signatures indeed correspond to intrinsic signalling activity, I should be able to see a higher pathway score in cancer patients with an oncogenic driver mutation

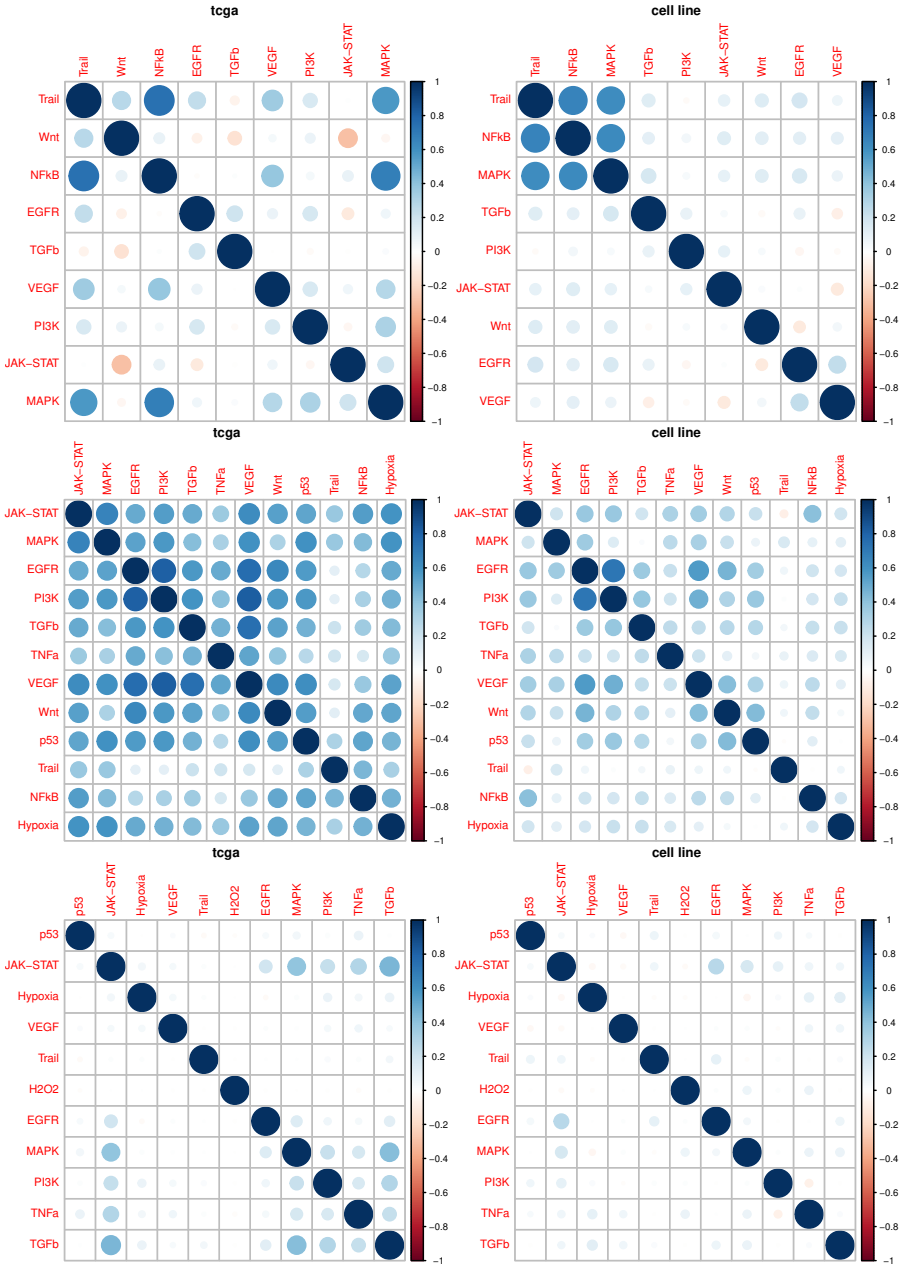

Figure 27: Comparison to state of the art pathway methods SPIA (top), Pathifier (middle) and PARADIGM (bottom). Legend as in figure 26.

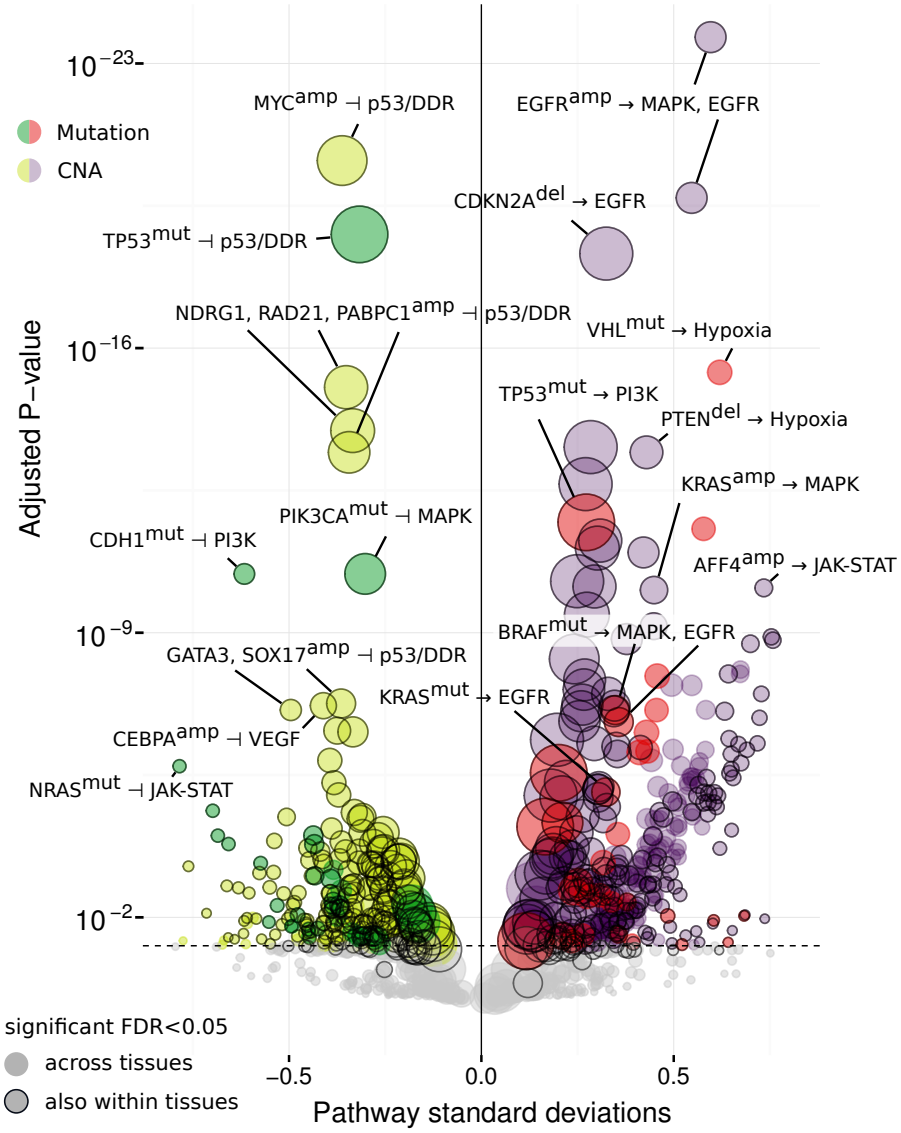

Figure 28: Volcano plot of associations between driver mutations (denoted by the superscript mut and green/red colour) and copy number aberrations (denoted by amp/del and yellow/purple colour) with inferred pathway activity using PRGs in basal gene expression of TCGA cohorts. Effect sizes larger than zero indicate inferred activation, smaller than zero inferred inhibition of a pathway. P-values FDR-adjusted with a significance threshold of 5%. Circles with a black outer ring are also significant if corrected for cancer type.

and a lower score when a tumour suppressor is mutated or lost compared to patients where no such alteration is present. Depending on how those aberrations are spread across cancer types, I should be able to detect them within or across cancer types.

I selected all cancer types in the TCGA for which there were tissue-matched normals available, in order to make full use of the pathway methods that require them. I calculated pathway scores for those using pathway-response genes, Reactome and Gene Ontology enrichment, SPIA, Pathifier, and PARADIGM. I used an ANOVA to calculate significant associations between the presence and absence of mutations and copy number alterations and the inferred pathway scores, both with and without regressing out cancer types (volcano plot in figure 28; other methods in figures B1 and B2, as well as tables in appendix B.1 and B.2 for mutations and CNAs, respectively).

In terms of proliferative signalling, I find that *EGFR* amplifications are correlated both with EGFR- and MAPK-responsive genes ($FDR < 10^{-20}$), and to a lesser extent PI3K, VEGF, and Hypoxia (FDR<10-9). *ERBB2* amplifications show an increase in EGFR and PI3K-responsive genes, but also a reduction in the Trail signature ($FDR < 0.05$), suggesting a relatively stronger impact on cell survival. *KRAS* mutations show an increase in inferred EGFR activity, and amplifications additionally for MAPK and PI3K ($FDR < 10^{-5}$). *BRAF* mutations have a positive effect on EGFR and MAPK ($FDR < 10^{-9}$) but not PI3K ($FDR > 0.4$).

For *TP53* mutations I find a significant reduction in p53/DDR activity ($FDR < 10^{-18}$) that also explains the associations with *MYC*, *RAD21*, *NDRG1*, and *PABPC1* ($p > 0.04$ if conditioned). It is also associated with a significant activation of the pathways for MAPK, PI3K, and Hypoxia ($FDR < 10^{-4}$). This is in contrast to loss of *TP53*, where I only find a reduction in p53/DDR ($FDR < 10^{-3}$) but no modification of any other pathway ($FDR > 0.15$). The dual nature of *TP53* mutations and CNAs are in line with the recent discovery that *TP53* mutations can act in an oncogenic manner in addition to the protein losing its tumour suppressor activity that has been shown for individual cancer types (Olive et al., 2004; C. Zhang et al., 2013; Weissmueller et al., 2014; Zhu et al., 2015). In addition, I can suggest a link between *TP53* mutations and genes that are specifically induced by activation of canonical oncogenic signalling.

I find that *VHL* mutations (which have a high overlap with Kidney Renal Carcinoma, KIRC) are associated with a stronger induction of hypoxic genes compared to other cancer types. More surprisingly, I find that presence of *PIK3CA* and *PIK3CB* amplifications and *PTEN* deletions is also more connected to increasing the hypoxic response ($FDR < 10^{-6}$) compared to an effect on the PI3K-responsive genes

(FDR between 10^{-2} and 10^{-5}). A role of PI3K signalling in hypoxia has been shown before (Zhou et al., 2004; Yang et al., 2009; Kilic-Eren, Boylu, and Tabor, 2013).

4.2.3 *Comparison of methods*

Comparing different methods (figure 29), I found that for *TP53* mutations (top left panel) only PRGs are able to recover the expected negative association between the mutation and p53/DDR activity. GO and Reactome showed a much weaker effect in the same, while Pathifier and SPIA showed an incorrect positive effect. For *KRAS* mutations (top right panel) only my method can detect a strong activation of the MAPK/EGFR pathways where the other methods either show no significant effect or an effect in the wrong direction. The same goes *EGFR* amplifications (bottom left panel). Across tissues, my method is the only one to recover hypoxia as the expected (Maxwell et al., 1999) strongest link with *VHL* mutations (bottom right panel).

4.3 CELL LINE DRUG RESPONSE USING THE GDSC

4.3.1 *Drug associations using GDSC cell lines*

The next question I tried to answer is how well our pathway-responsive genes are able to explain drug sensitivity in cancer cell lines. I took IC_{50} values from the GDSC project (Iorio, Knijnenburg, et al., 2016) and calculated statistical associations with our method, and GSEA using the same pathways in Reactome (Croft et al., 2011) or Gene Ontology (Gene Ontology Consortium, 2004), both for a pan-cancer and a tissue-specific condition.

I performed drug association using a linear model between 265 drug *IC50*s and 11 inferred pathway scores (`lm` function, R *stats* package), doing a total of 2915 comparisons for which I correct the p-values using the false discovery rate. For pan-cancer associations, I used the cancer type as a covariate in order to discard the effect that different tissues have on the observed drug response.

While this will also remove genuine differences in pathway activation between different cancer types, I would not be able to distinguish between those and other confounders that impact the sensitivity of a certain cell line from a given tissue to a drug. My pan-cancer association are thus correcting for different tissues when computing differences in drug response explained by inferred (our method, GO, or Reactome) pathway scores. For tissue-specific associations, I fit the linear model and correct p-values for the false discovery rate for each cancer type separately.

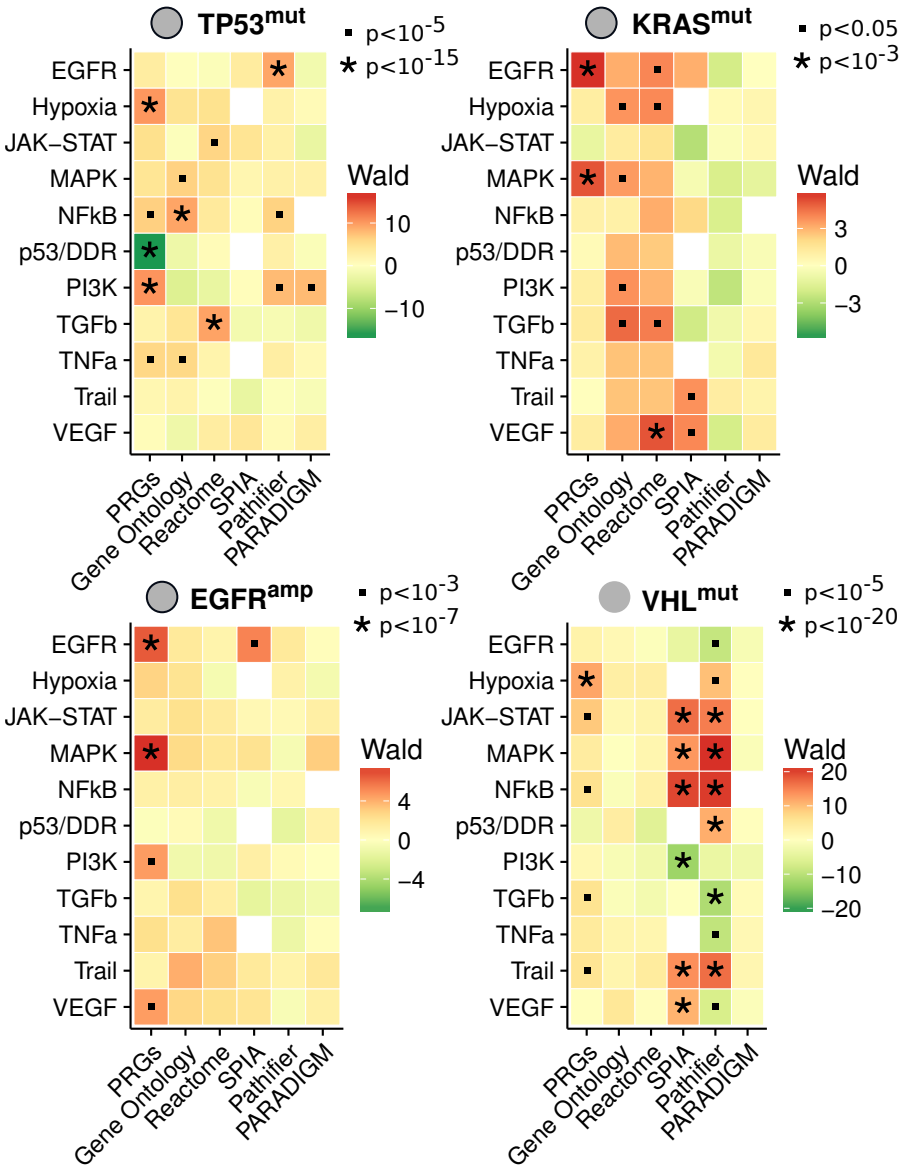

Figure 29: Comparison of pathway score (vertical axis) associations across different methods (horizontal axis). *TP*53 and *KRAS* mutations within cancer types in top row, *EGFR* amplification bottom left. *VHL* across cancer types. Wald statistic shown as shades of green for downregulated and red for upregulated pathways. P-value cutoffs shown as indicated. White squares if a pathway could not be used for a method.

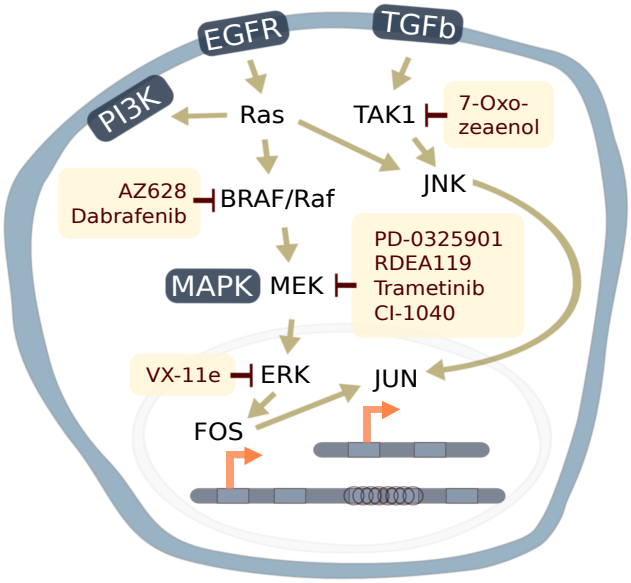

Figure 30: Pathway context of EGFR/MAPK and its inhibitors.

4.3.2 *Pan-cancer Associations with Drug Response*

I found that for the pan-cancer setting, there were 199 significant associations (FDR<10%, conditioned on tissue of origin) for my method and 27 and 66 for Gene Ontology and Reactome pathways, respectively. The top hit using my method was the associations between Nutlin-3a and p53-responsive genes. Nutlin-3a is an MDM2-inhibitor that in turn stabilizes p53, where it has also previously been shown that a mutation in *TP53* is strongly associated with increased resistance to Nutlin-3a (Garnett et al., 2012). This is thus a well-understood mechanism of sensitivity (presence) or resistance (absence of TP53 activity) to this drug that our method recovers but neither GO or Reactome pathways do.

I also find strong association between different MEK inhibitors (Trametinib, RDEA119, CI-1040, etc.) and MAPK/EGFR activation, but also a Raf (AZ628) or TAK1 (7-Oxozeaonol) inhibitor. These are all associations that the other methods miss or associate with a different pathway (figure B3 and tables in appendix B.3).

The other pathway methods showed a much lower number of associations across the range of significance (figure 32 and appendix B.3). Using the same significance threshold (10% FDR) mutated driver genes only yield 136 associations. They provide stronger associations only for *TP53*, where the signature is a compound of p53 signalling and DNA damage response, and PLX4720/Dabrafenib, drugs that were specifically designed to target mutated *BRAF*. For 170 out of 265 drugs covered by significant associations with either PRGs or mutations, the PRGs provided stronger associations for 85, with a signi-

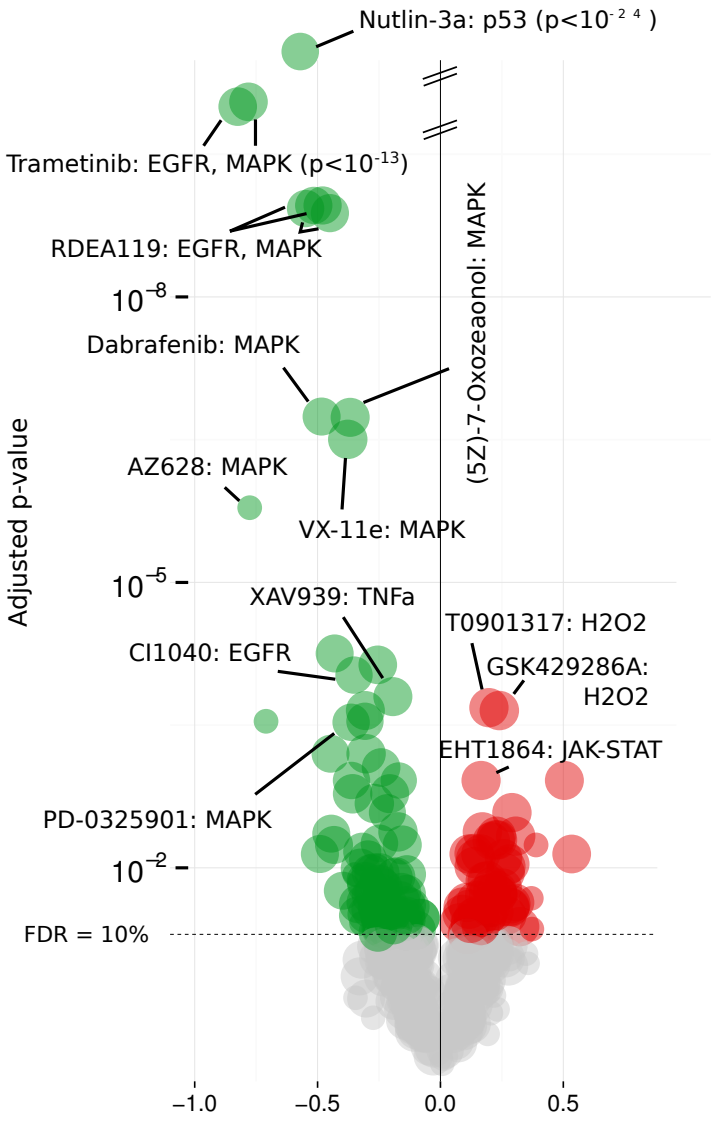

Figure 31: Volcano plot of pan-cancer associations between PRGs in GDSC cell lines and drug response (IC_{50}), corrected for cancer type. Effect sizes of smaller than zero indicate sensitivity markers (shown in green) and greater than zero resistance markers (shown in red). P-values FDR-corrected.

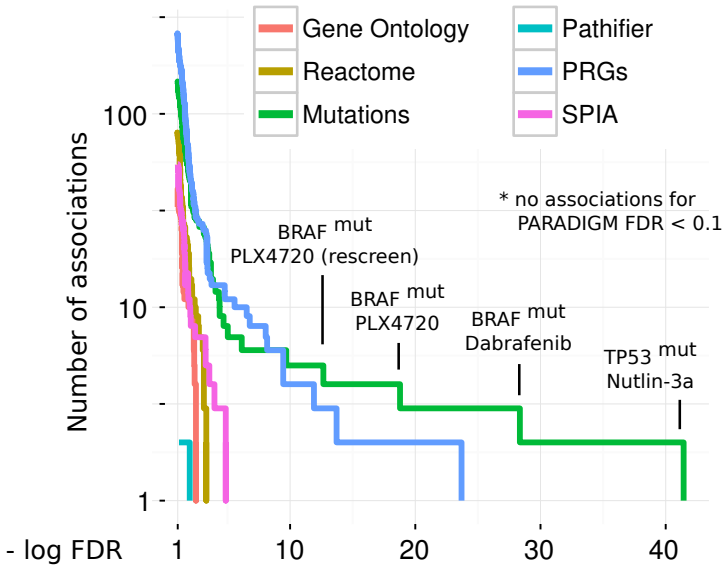

Figure 32: Comparison of the associations obtained by different pathway methods. Number of associations on the vertical, FDR on the horizontal axis. PRGs yield more and stronger associations compared to all other pathway methods. Mutation associations are only stronger for TP53/Nutlin-3a and drugs that were specifically designed to bind to a mutated protein. PARADIGM not shown because no associations < 10% FDR.

ficant enrichment in cytotoxic drugs compared to targeted drugs for mutations (Fisher's exact test, p<0.002).

However, stratification using PRGs and mutated driver genes is not mutually exclusive. My pathway scores are able to further stratify mutated and wild-type sub-populations into more and less sensitive cell lines (shown in figure 33 with additional statistics in table 10). This includes, but is not limited to, *BRAF*, *NRAS* or *KRAS* mutations using MAPK pathway activity and the MEK inhibitor Trametinib (top left) or Raf inhibitor AZ628 (bottom left), *BRAF* mutations with Dabrafenib (top right), and *TP53* mutations with p53/DDR and Nutlin-3a (bottom left). For MAPK- and *BRAF*-mutated cell lines, I find that cell lines with an active MAPK pathway according to the PRGs are 175 (AZ628), 7596 (Trametinib), or 105 fold (Dabrafenib) more sensitive than those where it is inactive. For Trametinib, cell lines with active MAPK but no mutation in *BRAF*, *NRAS* or *KRAS* are six times more sensitive than cell lines that harbour a mutation in any of them but MAPK is inactive.

Taken together, these results indicate that my pathway scores can be used to complement mutation-derived biomarkers by either refining them or providing an alternative where no such marker exists.

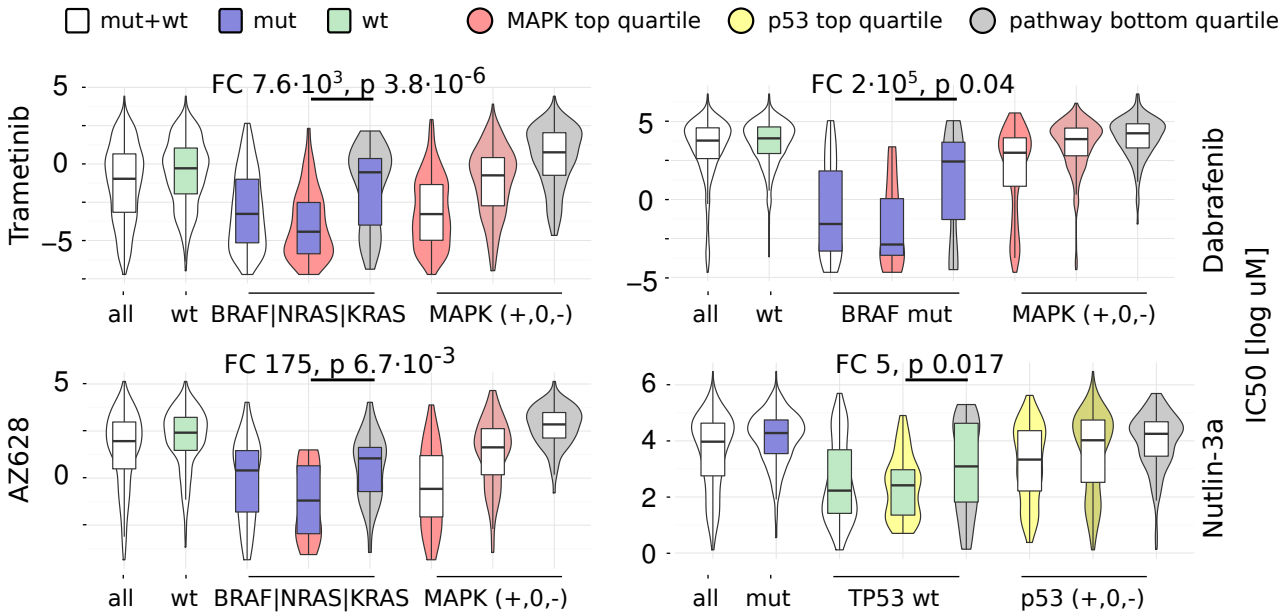

Figure 33: Comparison of stratification by mutations and pathway scores for the MAPK pathway (*BRAF*, *NRAS*, or *KRAS* mutations) and Trametinib (top left)/AZ628 (bottom left), *BRAF* mutation and Dabrafenib (top right), and p53 pathway/*TP*53 mutations/Nutlin-3a (bottom right). Mutations indicated by colour of the box (from left; blue for mutated, green wild-type, white mixed) and pathway scores by colour and shade of the violin (from right; pink for MAPK and and yellow for TP53). The more sensitive genomic phenotype (middle; mutated for MAPK/*BRAF* and wild-type for *TP*53) is further stratified by top- and bottom quartiles of pathway score (fold change of medians and p-value of Mann-Whitney U test as indicated).

Table 10: Stratification statistics: significance tests for figure 33. For the same subsets, results of the Mann- Whitney U test between different quartiles of the pathway score within different subsets defined by mutational status with p-value as indicated. Difference in mutations indicated by wt (wild-type), mut (mutated), or blank (any). Inferred pathway activity is indicated by + (top quartile) − (bottom quartile) or blank (any). Distance reported as a fold change (FC) of medians.

Treatment	Reference	Comparison	p-value	FC (medians)
MAPK + Trametinib	MAPK wt	MAPK mut	4.65e-29	950
MAPK + Trametinib	MAPK+	MAPK-	3.45e-33	10819
MAPK + Trametinib	MAPK+ wt	MAPK- wt	6.53e-13	166
MAPK + Trametinib	MAPK+ mut	MAPK- mut	3.83e-06	7596
MAPK + AZ628	MAPK wt	MAPK mut	1.03e-14	102
MAPK + AZ628	MAPK+	MAPK-	2.11e-14	2665
MAPK + AZ628	MAPK+ wt	MAPK- wt	7.69e-07	20
MAPK + AZ628	MAPK+ mut	MAPK- mut	6.72e-03	175
BRAF + Dabrafenib	BRAF wt	BRAF mut	1.88e-24	314085
BRAF + Dabrafenib	MAPK+	MAPK-	3.12e-13	18
BRAF + Dabrafenib	MAPK+ wt	MAPK- wt	1.93e-04	4
BRAF + Dabrafenib	MAPK+ mut	MAPK- mut	4.02e-02	212416
p53 + Nutlin-3a	*TP*53 wt	*TP*53 mut	3.09e-35	113
p53 + Nutlin-3a	p53+	p53-	9.60e-08	8
p53 + Nutlin-3a	p53+ wt	p53-	1.69e-02	5
p53 + Nutlin-3a	p53+ mut	p53- mut	9.29e-01	1

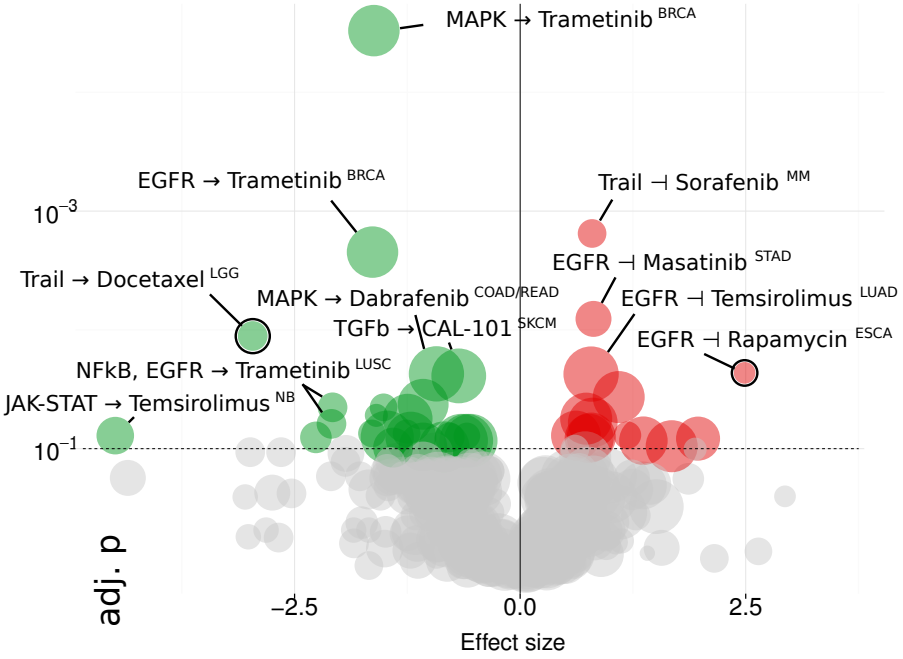

Figure 34: Volcano plot of tissue-specific associations between pathway activity and drug response for clinical drugs only. Horizontal axis shows the regression slope, vertical axis the per-tissue FDR-adjusted p-values. Dots with black circles are shown in more detail below.

4.3.3 *Tissue-specific Associations with Drug Response*

I performed drug associations also on the tissue-specific level. However, by dividing my dataset into the different cancer types I lose statistical power, so I chose in the first instance to only look at the subset of clinically approved drugs for the already stratified populations.

A drug association analysis for each tissue separately (figure 34) resulted in 75 significant associations for PRGs vs. 36 for Reactome and 12 for GO (regression, FDR<10%; associations of different methods in figure B4 and tables in appendix B.3). The strongest association was response to the MEK inhibitor Trametinib where sensitivity was positively correlated with expression of MAPK-responsive genes, explaining a total of 54% of the total variance observed in the IC_{50}s of BRCA drug response ($p < 10^{-8}$, $FDR < 10^{-5}$). My method is the only one able to recover this association, which is expected due to oncogene addiction to a hyper-activated MAPK pathway, or any other associations involving EGFR/MAPK for BRCA.

Perhaps more interestingly, I found two associations to drugs with strong effect sizes. The inferred activity of Trail correlates with increased sensitivity to Docetaxel in Low-Grade Glioma (LGG), where an increased response to Docetaxel (figure 35) in combination with a Bcl-2 inhibitor has been previously shown in BRCA (Lyseng-Williamson and

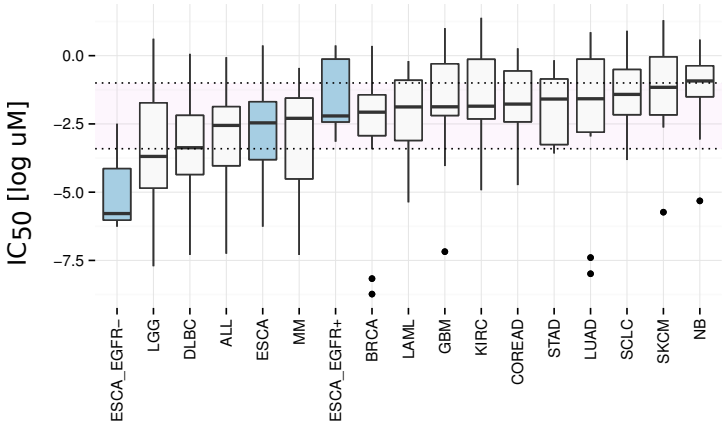

Figure 35: EGFR activity may mediate resistance to Rapamycin (macrolide antibiotic) in Esophageal Carcinoma (ESCA). Coloured boxes show the drug response of a certain tissue and the stratification achieved by negative and positive scores for a given pathway, white boxes indicate range of drug response for other tissues.

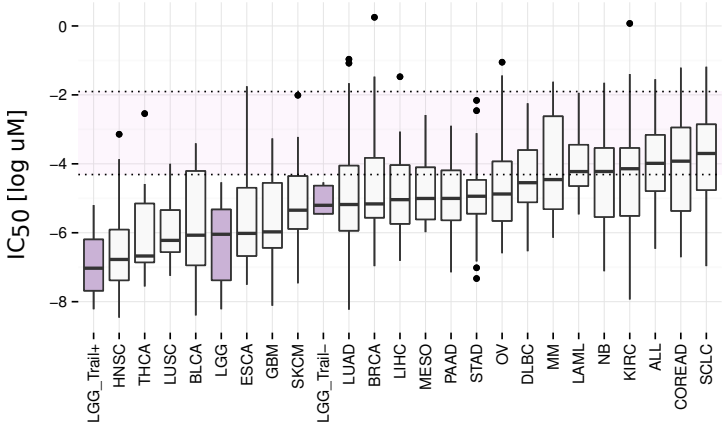

Figure 36: Trail activity mediates sensitivity to Docetaxel (microtubule stabilizer) in Low-Grade Glioma (LGG). Legend as in figure 35.

Fenton, 2005). Also, EGFR activity correlates with increased resistance in Esophageal Carcinoma (ESCA) cell lines treated with Rapamycin (figure 36). While BRCA and ESCA are already more sensitive to these drugs than most other tissues (unlike the case above of BRCA for Trametinib), the additional stratification by pathway scores identifies a subset comprised of half the cell lines per tissue with a more sensitive median response compared to all tissues, including those where the drug is in clinical use.

4.4 PATIENT SURVIVAL USING THE TCGA

The implications of inferred pathway activity compared to pathway expression is expected to be less clear for patient survival than for cell line drug response due to more heterogeneity and many more confounding factors involved that affect the phenotype observed. Nonetheless, I was interested in how my inferred pathway activity compared to the pathway expression methods in terms of overall patient survival. I would expect activity of canonical oncogenic pathways to be negatively correlated with patient survival and pro-apoptotic pathways to be positively correlated.

4.4.1 *Clinical data and methods*

Using the Firehose tool, I downloaded clinical data (file names including *Merge_Clinical.Level_1*, unpack them, select all *clin.merged.txt* files) for all cancer types for which it was available. I extracted the fields for study (`admin.disease_code`), age of the patient (`age_days`), their vital status (where a field `days_to_death` with a missing value indicated that the patient was alive, encoded with 0 if the patient is alive and 1 if they are not), days that they survived (either `patient.days_to_death` or if that was not available `patient.days_to_last_followup`), the TCGA patient barcode (`patient.bcr_patient_barcode`), and their sex (`patient.gender`). I converted the days of survival to months by dividing by 30.4. I discarded all patients that had a negative survival time. I removed duplicates, and in the case there were multiple records of the same patient I took the one recorded latest (the row where the field `patient.days_to_birth` had the highest number).

Starting from the pathway scores derived with GO/Reactome GSEA, SPIA, Pathifier, PARADIGM, and my method on the TCGA data as described above, I used Cox Proportional Hazard model (R package *survival*) to calculate survival associations for pan-cancer and each tissue-specific cohort. For the pan-cancer cohort, I regressed out the effect of the study and age of the patient, and fitted the model for each pathway and method used. For the tissue-specific cohorts, I regressed out the

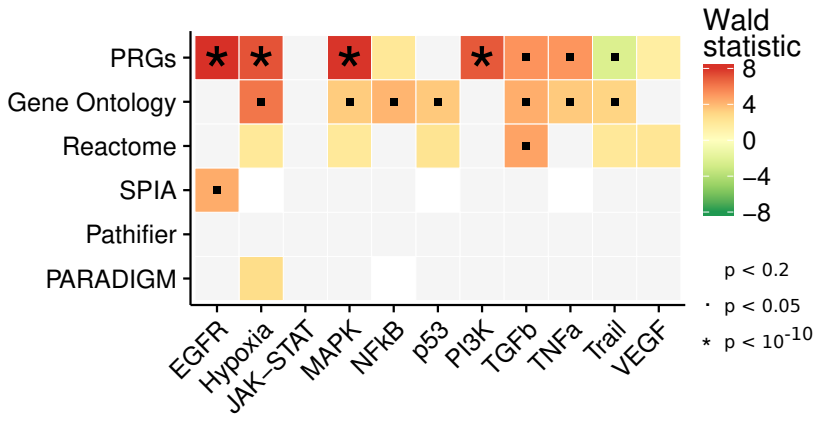

Figure 37: Comparison of pathway methods by association of a pathway with increased (green) or decreased (red) patient survival. Colours correspond to z-score (grey: FDR>20%, white: no pathway available). Method (vertical) and pathway (horizontal) labels as indicated. For our method, MAPK/EGFR provide the strongest negative contribution to survival. Trail/apoptosis is the only pathway mediating prolonged survival, but missing the significance threshold.

age of the patients. I adjusted the p-values using the FDR for each method and for each method and study separately. I selected a significance threshold of 5 and 10% for the pan-cancer and cancer-specific associations respectively.

4.4.2 *Pan-cancer associations with survival*

The pathway activity inferred by PRGs showed a strong association with decreased survival for EGFR, MAPK, PI3K, and Hypoxia (figure 37; associations of different methods in figure B5 and tables in appendix B.4). Gene Ontology found much weaker associations for those pathways, and the other methods missed them almost entirely. In terms of Trail activity, PRGs find an increase in survival while the other methods show either a decrease or no effect. For JAK-STAT, NFkB, p53, and VEGF there are no significant associations that are picked up by more than one method ($FDR < 0.05$). Compared to pathway scores, driver mutations only showed a significant decrease in survival for *TP*53 ($FDR < 0.03$ vs. $FDR > 0.2$) with a weaker effect size.

4.4.3 *Tissue-specific associations with survival*

For individual cancer types, I found a similar separation between oncogenic and tumour-suppressor pathways for the associations using PRGs (figure 38; associations of different methods in figure B6 and tables in appendix B.4) that other methods fail to provide. In addition, I found

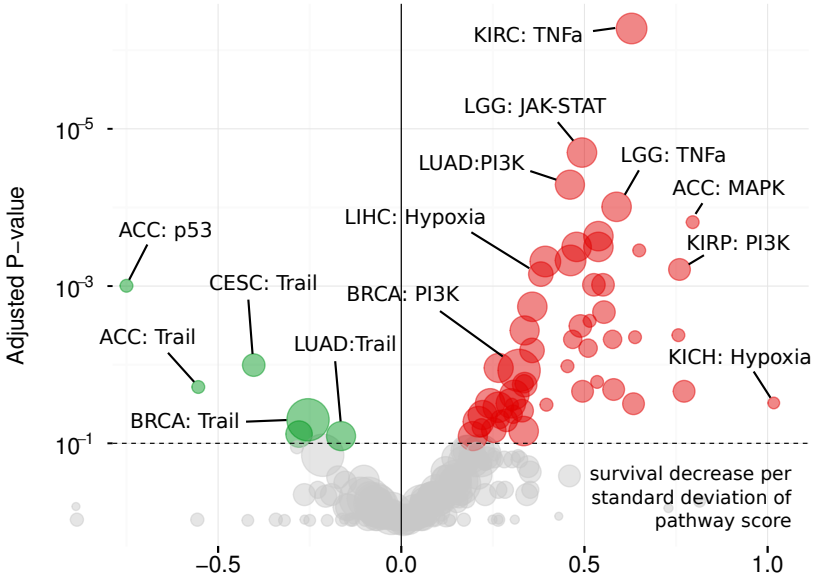

Figure 38: Volcano plot of cancers-specific associations between patient survival and inferred pathway score using our method, effect size on the x-axis, FDR-adjusted p-values on y. For these associations I can recover significant hits for apoptosis in multiple cancers and are the only method to do so.

cancer-specific associations of pathways with no effect in the pan-cancer setting. Adrenocortical Carcinoma (ACC) showed a significant survival increase with p53 activity ($FDR < 10^{-3}$). It is interesting that in this case also none of the samples harbours a reported gain-of-function TP53 variants (Zhu et al., 2015).

4.4.4 *Kaplan-Meier survival curves for specific hits*

For Kaplan-Meier survival curves (figure 39) it is only straightforward to plot discrete classes of input and not the continuous pathway scores I obtained using the different methods. In order to get distinct classes needed for interpretable survival curves, I split all obtained pathway scores in upper, the two middle, and lower quartile and respectively assigned the classes “up”, “unknown”, and “down” to show for the three examples of associations found.

As already found with the tissue-specific associations, Kidney Renal Clear Cell Carcinoma (KIRC) and Low-Grade Glioma (LGG) show decreased survival with TNFa and the latter with JAK-STAT, pathways where activating mutations are much less well established than for EGFR/MAPK.

For these associations, I found a difference in one-year survival of over 25% between the top and bottom quartiles of the assigned pathway scores (figure 39). Compared to mutations, PRG associations were

stronger (FDR 10^{-7} vs. 10^{-3}) and more consistent (strongest associations with small number of mutated genes).

4.5 DISCUSSION

The functional context in terms of cancer hallmarks that mutations create are similar despite obvious differences in the exact mode, as exemplified by the long tail of different variants seen in cancer genomes. For cancer diagnosis in both the clinical and preclinical setting, efforts like The Cancer Genome Atlas(The Cancer Genome Atlas Research Network et al., 2013) or the International Cancer Genome Consortium (ICGC) have pioneered this characterization on a large scale, offering the opportunity to derive large amounts of information about individual tumours and in turn likely harbours yet undiscovered treatment opportunities.

While the recent focus on linking outcomes with mutations represents a step forward over basing treatment of an individual on the tissue of tumour origin alone, putting mutations in a functional context with the signalling aberrations that they create may provide additional insight in mechanisms of pathogenesis and treatment opportunities.

In terms of drug associations, I have shown that my method outperforms the associations obtained by using GSEA on either pathway expression or GO modules, that strong tissue-specific associations are able to explain a significant part of the overall variability in the response to some drugs, and that pathway associations can be used to refine mutational biomarkers, or act as biomarkers themselves when there is no known associated mutation for a given drug.

I expect that this method of deriving signatures for pathway activity will also in other contexts be able to discover both oncogene addiction footprints to support available mutational biomarkers, as well as provide insights into pathway activation patterns that mediate sensitivity or resistance to a subset of drugs that other pathway-expression based methods are not able to.

For survival associations, only my signatures find the pathways that we would most expect to decrease patient survival by accelerating tumour growth (EGFR and MAPK) and promoting survival by apoptosis (Trail) to be associated with the respective outcome in both the pancancer and the tissue-specific cohorts. Other methods fail to separate those, only obtain significant associations for a very limited number of cancer types, and show high correlation between pathways.

Overall, my results suggest that consensus pathway response signatures provide a better measure of pathway activity than pathway expression, irrespective of whether the latter was derived from gene sets or directed paths. I have shown that they are able to refine our under-

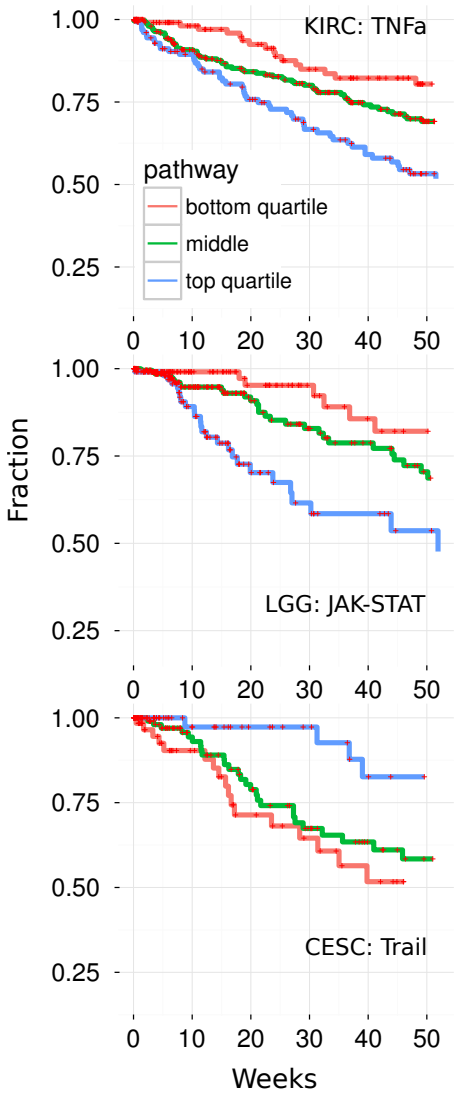

Figure 39: Examples of Kaplan-Meier curves where our method is able to separate between groups of up-, downregulated pathways vs the rest. Hits shown are TNFa in KIRC, JAK-STAT in LGG, and Trail and CESC, where the stratification between top and bottom quartile is always at or greater than 25% after a year.

standing of the impact of mutations, as well as their utility for cell line drug response and patient survival. The examples I outlined show that a downstream readout should be used as a proxy for pathway activity instead of mapping mRNA expression levels to signalling molecules.

5

MODELLING DRUG INTERACTIONS

At the time I started this project, the LINCS Connectivity Map (cf. section 1.5.5) data had just been released. The BROAD institute held a small symposium and workshop about the platform (the L1000), the data they had collected, and how they were planning to take it forward with web-based "apps" that did signature matching based on GSEA without actually exposing the data to its users.

Despite the drawbacks of the only 978 genes measured and the lack of a publication, it would be of enormous interest to combine this data set with what I am already working on with the GDSC: this way, I have for the same cancer tissue (sometimes even the same cell line) not only high-quality baseline expression and drug response curves, but also lower coverage/quality drug-perturbed gene expression for a lot of different compounds, including 150 drug that are overlapping between the GDSC and the LINCS project.

I could use these for signature matching methods. Even more interesting, I could generate a signature between cells that are sensitive and those that are resistant to a certain kind of drug, and then match this signature with another drug that potentially converts the cells from a resistant to a sensitive phenotype. If this interaction is causal, I would expect the two drugs to work better in combination than any single treatment suggests, hence providing a synergistic effect between the two.

This has been shown to work when trying to overcome glutocorticoid resistance in acute lymphoblastic leukaemia (ALL) (Wei et al., 2006), where the authors identified Rapamycin as a modulator of the resistance phenotype. Now, I could do it systematically with the 150x150 drugs overlapping between the LINCS and GDSC projects, and could thus obtain the first view of drug sensitisation over a large target space. This could ultimately lead to a new drug combination or regimen that restrict the development of resistance in clinical cancer patients, which is one of the main challenges in patient treatment today (Ramaswamy, 2007).

Results of this chapter are unpublished. I produced all analyses, plots, and written text in this thesis myself. Data for experimental validation in chapter 5.3 was produced by and obtained from Ole Pless, Bernhard Ellinger, and Milena Kalmer (Department ScreeningPort, Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Hamburg).

5.1 MANTRA AND THE ORIGINAL CONNECTIVITY MAP

5.1.1 *Two-Tailed GSEA using MANTRA*

The original Connectivity Map (Lamb et al., 2006) consisted of 6100 perturbation experiments on five different human cell lines (but mostly the MCF-7 breast cancer cell line) using 1309 different, mostly non-cancer-related, small molecules as perturbations and microarray data as readout. On top of this data set (Iorio, Bosotti, et al., 2010) ordered differentially expressed genes by their fold change for each drug, and then merged those inversely weighted by distance[1] to arrive at what they call a Prototype Ranked List (PRL) that corresponds to a consensus signature across different conditions for a given drug–using a method called Mode of Action by Network Analysis (MANTRA).

The microarrays of the Connectivity Map (HG-U133A) and the GDSC (HG-U219) only had 83 overlapping probe set identifiers, so I mapped the probe sets in the PRLs to HGNC symbols. From the 250 most up- and downregulated genes in the PRLs I took those that are uniquely present in the respective subset, leaving approximately 100 genes per drug signature. I performed a two-tailed GSVA (details section 2.2.2) in the basal expression of the GDSC cell lines and thus obtained a drug-responsive signature expression score for each drug that has a PRL and each cell line in the GDSC panel. The final enrichment score was composed of the two individual enrichment scores of of both parts of the list, using:

$$ES = ES^{up} - ES^{down}$$

In comparison to GSVA, GSEA scores and signature-scaled GSEA scores are bimodally distributed around zero (cf. section 2.1.2). One could argue that with the up- and downregulated parts of the signature, a combination of the two would yield a unimodally distributed score again and thus GSEA scores can be used in conjunction with tests that assume a normal distribution of scores. This is not true because scores that are above zero with one signature can be either above or below zero in the other signature. In practice, combining GSEA scores yields a trimodal distribution (figure 40, left panel). In contrast, GSVA scores are normally distributed for each individual score as well as for the combined score (figure 40, right panel), which is why I chose GSVA over GSEA also in this context.

1 so different conditions are weighted equally and over-representation of one condition, like the experiments performed in the MCF-7 cell line, do not bias the data set

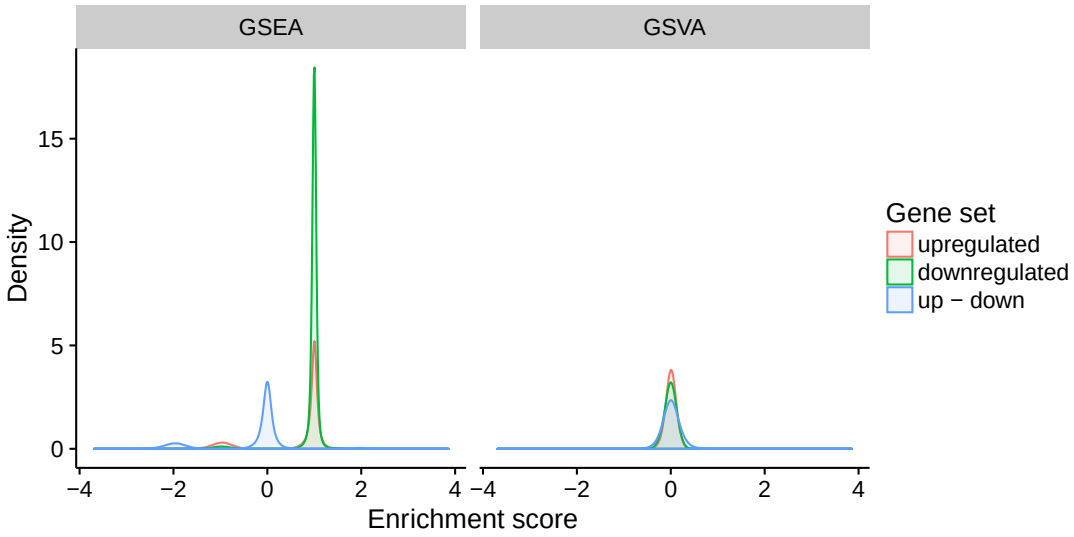

Figure 40: Distributions of GSEA (left panel) and GSVA (right panel) scores for up- and downregulated genes as well as the combined score of both. GSEA: Combining up- and downregulated score yields a distribution around zero (if ES^{up} and ES^{down} have the same sign) or more extreme tails if they have opposite signs (e.g. an ES^{up} of -1 and ES^{down} of 1 would yield a combined score of -2, like the peak at that position in the left panel; there are also scores distributed around 2, but their peak is barely visible). GSVA: Both the individual and the combined score are unimodally distributed and resembling a normal distribution, making them more amenable to statistical testing.

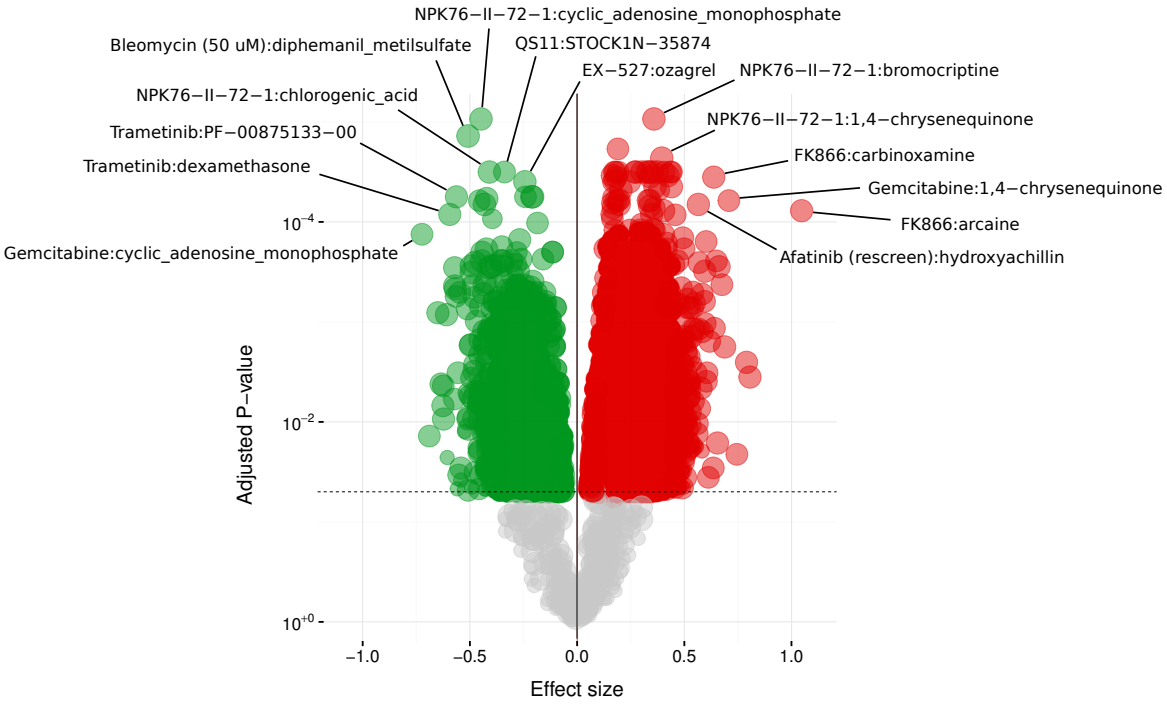

Figure 41: Volcano plot of associations between expression of drug signatures (second part of label) and drug response (first part) for cell lines of all tissues. Effect size is standard deviations of the score on the horizontal axis, FDR-adjusted p-values on the vertical axis.

5.1.2 *Pan-cancer associations*

I calculated associations as described in section 2.1.3. Results of associations between the expression of a drug-response signature (the sensitiser) and drug sensitivity to a second drug (the treatment drug) are shown as volcano plot in figure 41 (associations in appendix section C.1). The top hits mostly include the drugs NPK76−II−72−1 (an experimental PLK3 inhibitor), Trametinib (a MEK inhibitor) Afatinib (an EGFR inhibitor), and Bleomycin/Gemcitabine (cytotoxic drugs). The general-purpose drugs in the Connectivity Map that induce gene expression changes correlated with increased sensitivity in those are for instance cAMP, chlorogenic acid (antioxidant and involved in the lignin biosynthesis), bromocriptine (a dopamine promoter), or 1,4−chrysenequinone (a DNA-binding metallo-intercalator). It is not straightforward to propose a possible mechanism in which each of the two drugs interact. Furthermore, the concentration of all drugs in the Connectivity Map is 10 uM, which is much higher than the actual bodily concentration in most cases. I have no indication that the associations I found should indeed be causal and act synergistically or antagonistically *in vivo*. It would be much more interesting to look at combinations of two cancer drugs or at least cancer-related drugs. In that regard, the most inter-

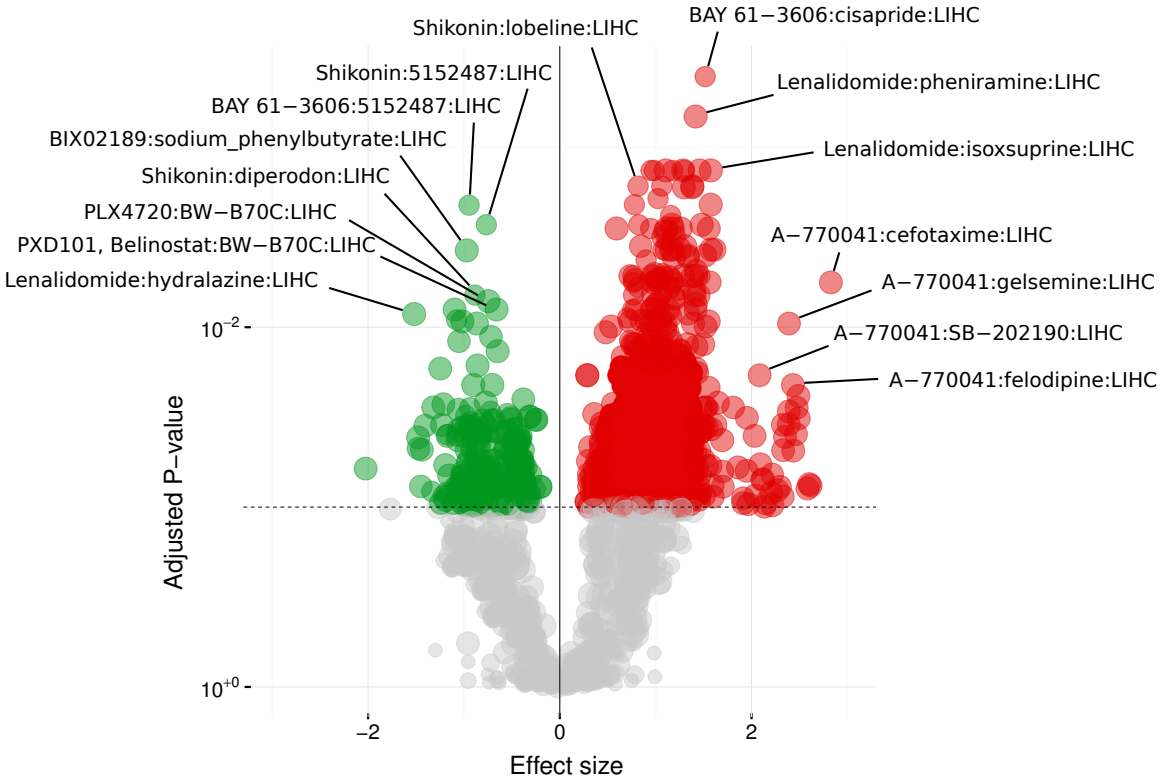

Figure 42: Volcano plot of associations between expression of drug signatures and drug response for individual cancer types. Effect size is standard deviations of the score on the horizontal axis, FDR-adjusted p-values on the vertical axis.

esting associations is Trametinib with dexamethasone, as the latter is administered with chemotherapy already.

5.1.3 *Tissue-specific associations*

In addition to the potential issues outlined above, tissue-specific associations are a lot less stable due to lower sample numbers. In my case, they also (figure 42 and associations in appendix section C.1) yield all the strongest hits with Liver Hepatocellular Carcinoma (LIHC), for which there is no reason to assume that *a priori*. While these associations could hint at possible synergistic or antagonistic combinations, I can not easily make sense of them. One of the strongest associations, for instance, is a drug targeted at BRAF (PLX4720), a mutation that does usually not occur in liver cancer. After discussions with our collaborators at the Sanger, we decided that these combinations are too far from available biological knowledge to test them further.

5.2 A PAN-CANCER VIEW OF DRUG SENSITISATION USING LINCS

5.2.1 *Data quality of the LINCS*

The fact that the paper about the LINCS data has not yet been published, combined with the concerns raised about data quality makes it necessary to point out that in fact we do not yet know its use and limits of applicability. While a comprehensive evaluation of the data is beyond the score of this thesis, it is important to perform quality control measurements to make sure the changes in gene expression reported via z-scores are indeed biologically meaningful. I did this in the following ways: (1) generate the same signatures as in chapter 3 and see how well they agree on drug associations, (2) for each drug perturbation I look at here quantify how well its signature is able to recover the same perturbation compared to other perturbations. Together, they provide evidence that the data is noisy but usable for the purpose of signature matching.

In terms of oncogene addiction associations, I find very similar results to the ones obtained in section 4.3.2. The most significant association is the p53 pathway with Nutlin-3a, followed by MAPK activity and MEK (and related) inhibitors. However, even with many more experiments used to generate the signature on either the 978 landmark genes or the projected set, the associations I obtain are a lot weaker ($p < 10^{-5}$ instead of $p < 10^{-20}$) for the strongest and all following associations.

In terms of how well signatures can recover the same drug that was used to generate them, I scored all input experiments with with the signature I obtained and quantified the area under the Receiver Operator (ROC) curve. I discarded all drugs that did not yield an AUC of > 0.7. This threshold is not very stringent and might still include drugs that have a relatively weak signature. However, we will only know whether this was a good cutoff once the proposed experiments in the later sections have been validated.

5.2.2 *Creating drug signatures and scoring GDSC cell lines*

What would be far more interesting is to investigate how drugs specifically designed or used in cancer interact in terms of synergies or antagonism. With the new LINCS Connectivity Map, that has become possible. There is a total of 150 drugs that are both in the GDSC and the LINCS, enabling me to do a comprehensive characterisation of signature matching. I derive models for each drug separately, with a linear model of LINCS-provided projected z-scores mapped to HGNC symbols in experiments where the drug was used vs. all others. I selected the top N genes as the signature (from 0.01 up to 10% FDR),

and chose N to maximise the area under the precision-recall curve of the current vs. all other perturbations. I kept all drug signatures that have an AUC of over 0.7 recovering the experiments where the same drug as I used and discard the rest. For the pan-cancer cohort, this leaves 123 out of 150 signatures.

Analogous to the model of pathway-responsive genes (chapter 3, implementation in section 3.2.2), I calculated the expression level of a drug-responsive signature using the linear transformation:

$$S = E * Z$$

Where E the basal gene expression values of the GDSC cell lines with cell lines in rows and genes in columns, Z are the z-scores of the fitted model with genes in rows and drugs in columns, and S the resulting expression of drug-response signatures of the GDSC cell lines with cell lines in rows and drugs in columns. I scaled S for each drug to have mean 0 and standard deviation 1.

5.2.3 *Naive associations with drug response*

For the pan-cancer drug associations, I check for the correlation between drug sensitivity and signature expression using the following linear model:

$$D_i \sim T + S_j \qquad \forall i \in GDSC\,drugs\ \forall j \in LINCS\,signatures$$

Where S_j is the expression of the signature in response to drug j across all cell lines, and D_i is the sensitivity towards drug i across all cell lines. T corresponds to the tissue of each cell line and is used to regress out the difference in drug response between tissues. I compute those associations for all sensitisers ($j \in LINCS$) and treatments ($i \in GDSC$), adjusting the resulting p-values by the false discovery method. However, note that because all signature scores are computed on the GDSC panel the LINCS signatures are used to provide scores for all GDSC cell lines. Hence, the cell lines indexed for building the model are all of the GDSC:

$$\begin{matrix} D_i^{c_1} \\ D_i^{c_2} \\ D_i^{c_3} \\ \vdots \end{matrix} \sim \begin{matrix} T^{c_1} \\ T^{c_2} \\ T^{c_3} \\ \vdots \end{matrix} + \begin{matrix} S_j^{c_1} \\ S_j^{c_2} \\ S_j^{c_3} \\ \vdots \end{matrix}$$

An overview of significant hits is shown in figure 43 (associations in appendix section C.2). The strongest potentially synergistic hit I obtain is Trametinib (GSK1120212) to act as a sensitiser for QS11

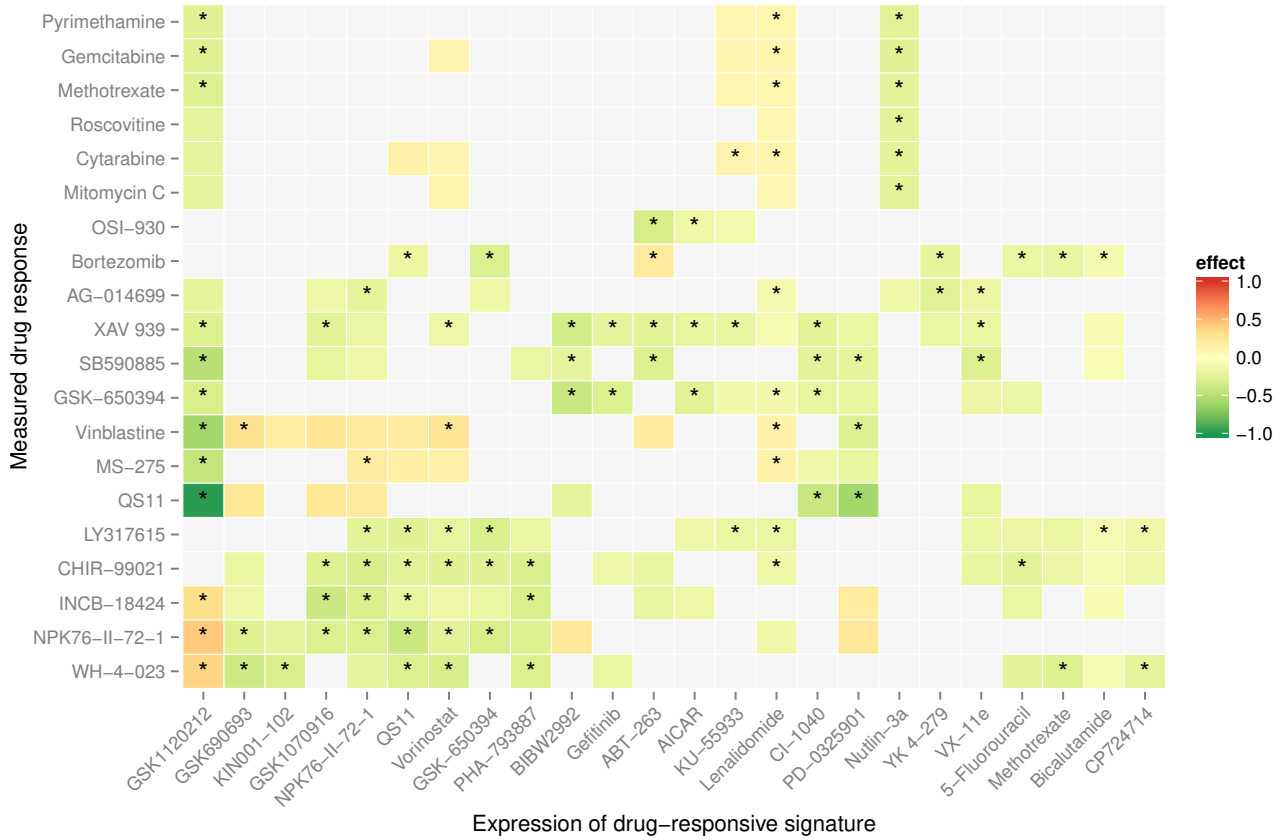

Figure 43: Matrix of naive drug sensitisation associations. Predicted synergies in green, antagonisms in red. Effect size corresponds to the regression slope and is shown if $FDR < 0.2$ and marked with * if $FDR < 0.05$. Selection of drugs on both axes chosen to show the subset with most synergies. The most significant synergistic hit is GSK1120212 (Trametinib; a MEK inhibitor) with QS-11 (a Wnt agonist) with $FDR < 10^{-5}$.

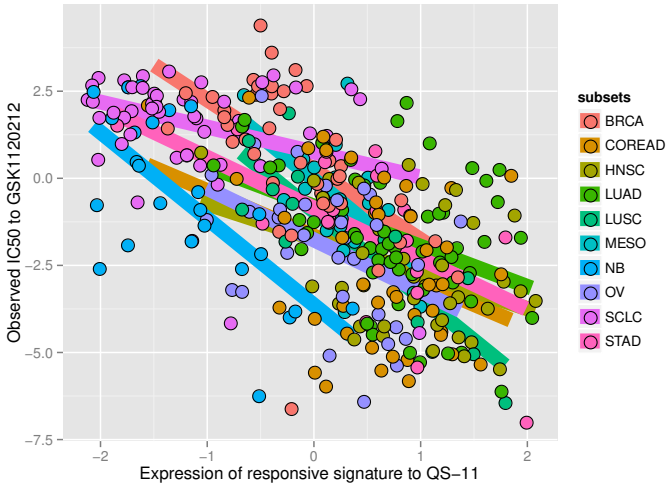

Figure 44: Linear fit behind the associations between Trametinib and QS-11. There is a strong linear correlation between expression of the QS-11-responsive signature and the drug sensitivity to Trametinib across all cancer types (figure 43), as well as a significant correlation for the cancer types shown in this figure $FDR < 0.05$.

($FDR < 10^{-5}$). The most significant hit overall is Trametinib that is antagonistic with itself ($FDR < 10^{-7}$; not shown in plot because it is focussed on the most synergistic examples). While the latter can not be true according to the definition of synergy and antagonism, the former provides a workable hypothesis and a strong fit across multiple tissues (figure 44). I could, however, not find any literature evidence to support it.

5.2.4 *Pathway correlation as a major source for false positives*

Setting aside the issue on whether an apparent lack of literature support can be a false positive or a new finding, an antagonistic interaction of a compound with itself goes against the definition of synergy (details section 5.3.2). This provides support for the hypothesis that the synergistic interaction between a MEK1/2 inhibitor and a Wnt agonist might be a false positive as well. So what mechanism could explain a strong association under the assumption that the effect is indeed a false positive? One indication is that Wnt activation is known to cross-activate the MAP kinase pathway (Smit et al., 2004).

From my results above, it seems that MEK inhibitors in general produce good signatures, or otherwise I would not expect as many MEK inhibitors as I see in my associations (GSK1120212, CI-1040, PD-0325901 in figure 43). A possible explanation is that those two effects together produce a QS-11 signature that is correlated with MEK or MAPK activation. MEK inhibitors will correspond to shutting off MEK

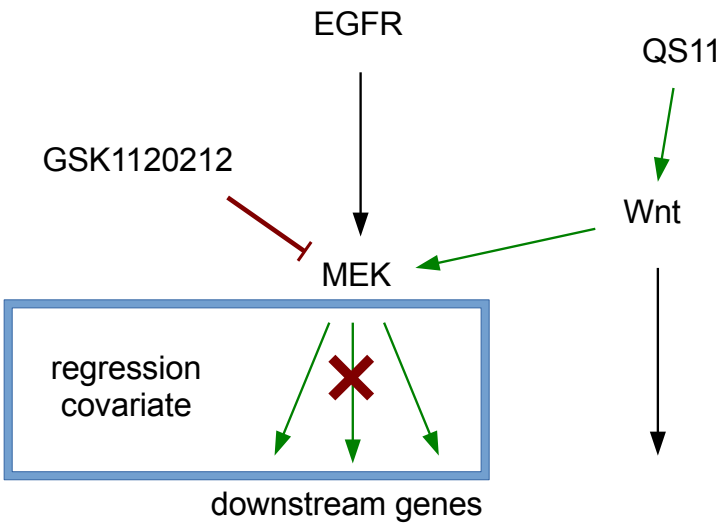

Figure 45: Potential reason for the association between the QS11-responsive signature and the MEK inhibitor Trametinib (GSK1120212): QS11 activates Wnt, which cross-activates MAPK. However, the gene expression signature of MAPK is stronger and thus the primary readout. Expression of this signature is thus correlated with MAPK activation, which is in turn correlated with increased sensitivity to MEK inhibitors.

(and the MAPK pathway), potentially reducing the associations to a correlation between pathway activation and inactivation that is strong statistically, but without real biological significance (schema outlined in figure 45).

One way to test the above hypothesis is to introduce as a covariate in the regression the expression score of the signature of the treatment drug, and calculate the associations as everything that is left after correcting for the effect on gene expression that it would have (using the expression score of Trametinib/GSK1120212 and calculating the association strength of the QS-11 signature and Trametinib response on top of that, as indicated with the blue box in figure 45). This, in more general terms, would correct for pathway cross-talk between the sensitiser and the treatment drug, but also for two-drug pairs that have the same target or are indeed the same drug. If the above association stays significant despite the added covariate, I can reject the model proposed in figure 45. Otherwise, this would be a valid explanation of the effect I see.

5.2.5 *Building an improved model*

As I am only looking at the drugs that are present in both the GDSC and LINCS data sets, introducing the signature of the first drug as a covariate is straightforward:

$$D_i \sim T + S_i + S_j \qquad \forall i \in \mathit{GDSC\,drugs}\ \forall j \in \mathit{LINCS\,signatures}$$

Where the addition of S_i indicates the conditioning on the drug-responsive signature of drug i. Following the model fit, I discard all coefficients of tissue or covariate signature, only looking at the results between D_i and S_j. Here again, the cell line c indexes GDSC cell lines and their expression of the LINCS-derived signatures:

$$\begin{matrix} D_i^{c_1} \\ D_i^{c_2} \\ D_i^{c_3} \\ \vdots \end{matrix} \sim \begin{matrix} T^{c_1} \\ T^{c_2} \\ T^{c_3} \\ \vdots \end{matrix} + \begin{matrix} S_i^{c_1} \\ S_i^{c_2} \\ S_i^{c_3} \\ \vdots \end{matrix} + \begin{matrix} S_j^{c_1} \\ S_j^{c_2} \\ S_j^{c_3} \\ \vdots \end{matrix}$$

This takes the significance of the Trametinib-QS11 pair away, as shown in the resulting association matrix in figure 46 (associations in appendix section C.2). Top hits now are Trametinib (GSK1120212) with PD-0332991 (a CDK4/6 inhibitor), FK866 (NAMPT inhibitor) with Sorafenib (a Raf inhibitor) and NPK76-II-72-I (NAMPT inhibitor). The strongest antagonistic hit is Thapsigargin (SERCA inhib-

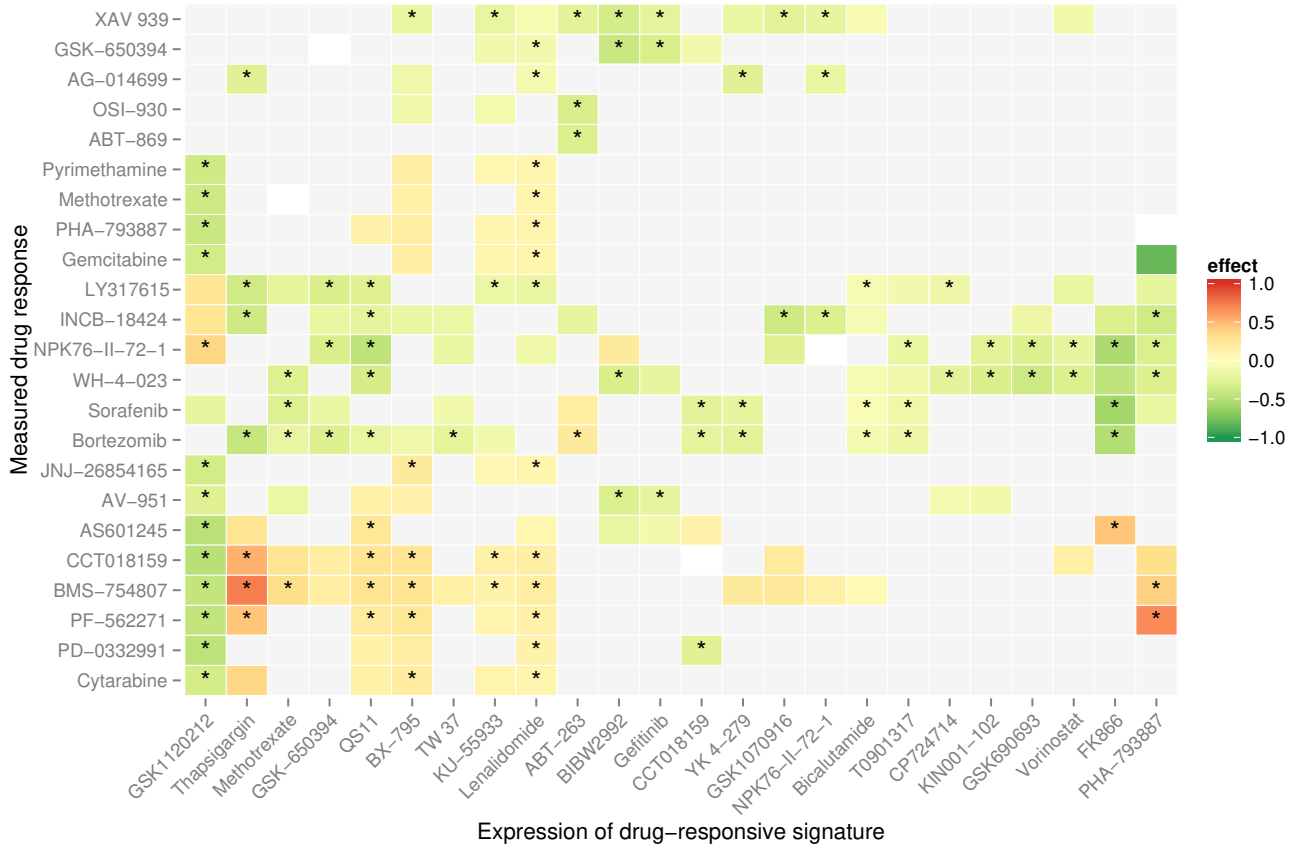

Figure 46: Matrix of associations including the treatment drug as covariate

itor) with BMS-754807 (IGF inhibitor). This approach also no longer yields associations of a drug with itself.

When looking for literature evidence to support some of the stronger hits in this matrix, I now do find some (figure 47):

- CDK inhibitors may act as sensitisers for Trametinib (GSK1120212) in multiple tissues (top). This has been found in a mouse model of melanoma (Kwong et al., 2012), my analysis suggests stronger effects in the tissues listed.

- Combination of Raf (SB590885) and MEK inhibitors in BRCA (middle). A synergistic effect has been shown in melanoma (Killock, 2014).

- Various drugs may act as sensitisers for Temozolomide (DNA alkylating agent) in COREAD (bottom). This has been shown for Mitomycin C and Rucaparib (PF-01367338, AG014699) in metastatic melanoma (Plummer, C. Jones, et al., 2008; Plummer, Lorigan, et al., 2013). Polo-like kinase 1 inhibition (target of GW843682X) has been shown to cause decreased proliferation

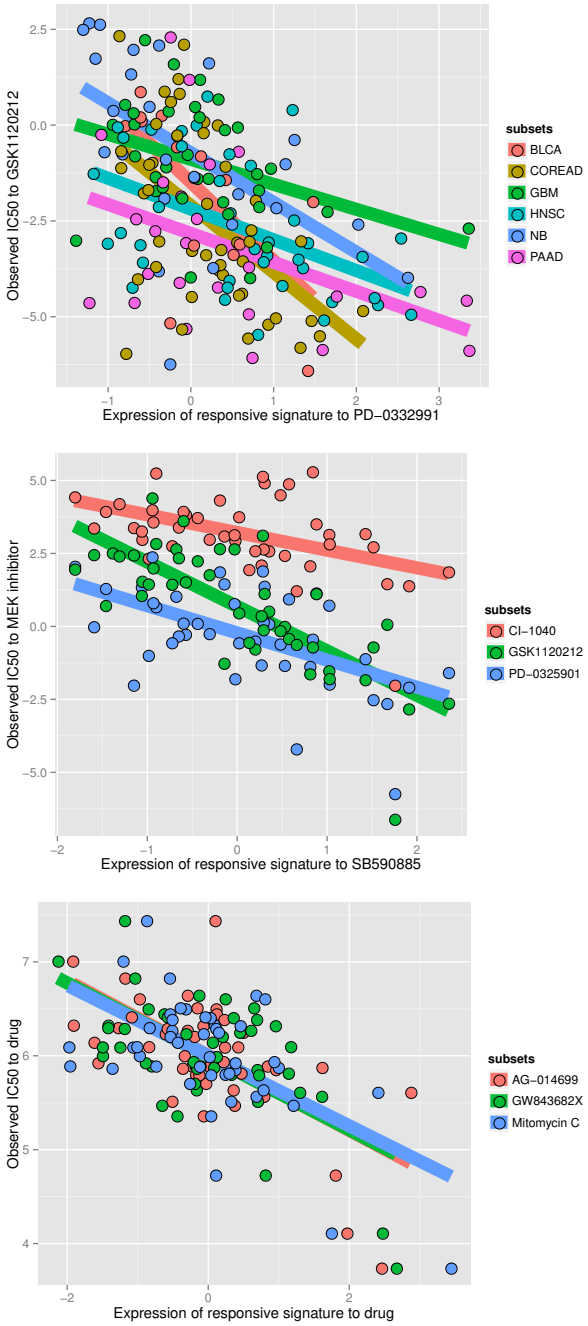

Figure 47: Linear fits behind the associations of PD-0332991 (a CDK4/6 inhibitor) and Trametinib (GSK1120212, a MEK inhibitor) in multiple tissues (top), the Raf inhibitor SB590885 and multiple MEK inhibitors in breast cancer (middle), and Temozolomide and various drugs (bottom). All associations FDR<0.05.

Figure 48: Overview of number of experiments available

by cell cycle arrest, leading to cell death in glioblastoma (Pezuk et al., 2013).

5.3 TISSUE-SPECIFIC SYNERGISTIC COMPOUNDS

5.3.1 *Consensus models for breast cancer cell lines*

While pan-cancer associations can uncover far-reaching implications of treatment with different drugs to make inferences about the effect of their combination, experimental follow-up on those must be more restricted in order to suggest a feasible setup with clear hypotheses. This is where tissue- or cancer-specific models come in: using signature matching, I derived hypotheses of potential synergies or antagonisms that can actually be tested in one or a few cell lines. Compared to pan-cancer associations, I would expect them to be more likely to be true for this particular cell line or small set of cell lines.

However, tissue-specific models of drug synergies are tricky. First, there is fewer cell lines to define the signature. A mutation that changes the gene expression response upon drug treatment is likely not going to have a large effect in the pan-cancer context because of the overall number of cell lines available, but if there is one in a certain tissue for which I derive a signature this may or may not change the outcome. Second, there are fewer cell lines with drug sensitivity data and thus the associations are much closer to the significance threshold.

In addition, there is a lot of choices to make about the model: Which tissue or which cell line to choose? Limit the drugs to a set where the tissue is sensitive in? How many cell lines are needed to derive meaningful signatures? Are the landmarks or projected genes better to make those inferences? Use a covariate to correct for pathway cross-talk or does it take away too much statistical power? These are all questions that could be answered in iterative cycle of predictions and experimental readout. Unfortunately, I can not do this.

I decided to focus on breast cancer because the number of available perturbations with cancer-relevant drugs far exceeded any other tissues (figure 48). But instead of taking the top-performing combinations

Table 11: Selected drug combinations for breast cancer cell lines

Sensitiser (24+72 hours)		Treatment (72 hours)		2*Combined score
Drug	Target	Drug	Target	
WZ3105	DDR1 RTK (?)	PD-0325901	MEK1/2	-3.14
AUY922	HSP90	AZD7762	Chk	-2.85
AT-7519	CDK	PD-0325901	MEK1/2	-2.66
AUY922	HSP90	ZM-447439	AURK	-2.64
PAC-1	Apoptosis inducer	AZD7762	Chk	-2.22
WZ3105	DDR1 RTK (?)	Trametinib	MEK1/2	-1.70
VX-680	AURK	AZD7762	Chk	-1.66
AUY922	HSP90	PD-0325901	MEK1/2	-1.65
AUY922	HSP90	BX-795	PDK-1	-1.56
AUY922	HSP90	Trametinib	MEK1/2	-1.48
WZ3105	DDR1 RTK (?)	BX-795	PDK-1	-1.44
PAC-1	Apoptosis inducer	ZM-447439	AURK	-1.18
VX-680	AURK	Trametinib	MEK1/2	-1.06
PAC-1	Apoptosis inducer	BX-795	PDK-1	-1.01
AUY922	HSP90	VX-680	AURK	-0.99
WZ3105	DDR1 RTK (?)	AZD7762	Chk	-0.88
VX-680	AURK	PD-0325901	MEK1/2	-0.82
PD-0325901	MEK1/2	VX-680	AURK	-0.81
AT-7519	CDK	Trametinib	MEK1/2	-0.72
PD-0325901	MEK1/2	AT-7519	CDK	-0.64
ZM-447439	AURK	PD-0325901	MEK1/2	-0.56
PAC-1	Apoptosis inducer	VX-680	AURK	-0.51
PD-0325901	MEK1/2	WZ3105	DDR1 RTK (?)	-0.50

from individual predictions, I would instead look at the ones that are most consistently predicted to be synergistic. I built models (analogous to section 5.2.2) for the individual cell lines MFC-7, SKBR-3, BT20, MDA-MD-231, HS-578T, MCF-10A, and the combined signature for breast cancer-related cell lines, for all projections, with and without covariate. I assigned a combined score by summing the logarithm of the adjusted p-value for each 2-drug combination across all different conditions. I expected that real synergistic combinations would systematically show up as more synergistic than neutral or antagonistic combinations.

I used a linear optimisation strategy (Integer Linear Programming, ILP) to select up to 25 combinations of 10 drugs that would maximise the overall synergistic significance while minimising antagonistic significance. I decided to exclude combinations of MEK and EGFR inhibition because that is already known. The result of this selection is listed in table 11.

5.3.2 *Defining synergy for drug combinations*

Models of synergy

Before I go on and try some of the predicted combinations, I first need to decide on how I expect two drugs to act in combination if there is no interaction effect, and then to quantify the interaction effect if there is one. Fortunately, there are common methods available to calculate synergy, the best-known of which are Highest Single Agent, Bliss Independence and Loewe Additivity.

Highest Single Agent (Laska, Meisner, and Siegel, 1994) is not really a synergy method at all. As the name suggests, this method takes the best-performing single agent at a given concentration and uses this as a baseline upon which to define synergy if a second drug improves upon the single treatment or antagonism if the combination has a weaker effect than any single agent. This, however, labels all additive effects as well as the interaction between a drug and itself as synergistic.

The Bliss Independence model (Greco, Bravo, and Parsons, 1995) makes a more sensible assumption: if both drugs don't interact, they exert an effect independently of each other. Thus, if one drug kills off half of the cells in the well, the other can be seen as acting only on the remaining half. Synergy and antagonism are defined as an effect above and below that, respectively. This model, however, may still label a drug to act synergistically with itself (Zhao et al., 2014): if you consider the single-agent dose response curve of AT-7519 in figure 50 and apply the drug at its IC_{50}, adding the same amount of the same drug again would already put it above the IC_{75} (use the red dots for suggested combination dilutions as a guide). This is more likely to happen the steeper the dose-response curves are.

The Loewe Additivity model (Loewe, 1926; Greco, Bravo, and Parsons, 1995) assumes the null interaction to be as if one drug is a dilution of the other. If we consider this assumption stringently, it means that the model should be used if two drugs share the same mechanism of action. Nevertheless, it is not prone to errors as the ones above and was the model of choice for the DREAM drug synergy prediction challenge.[2]

A model based on Loewe additivity

Given that we know which concentration x of a given drug we treat our cells with and which relative (to the untreated control) viability y we observe, we can calculate the Hill parameters for the minimal viability E^{min} (at high drug concentration), the maximal viability E^{max} (at very low drug concentration), the concentration of the half maximum effect m and the slope of the sigmoid λ. Note that E^{min} and E^{max} are

2 https://www.synapse.org/#!Synapse:syn4231880/wiki/235645

usually for the minimum and maximum effect instead of relative viability. I can use them for viability here because both range between 0 and 1, the only difference is that my curve is that in this case I am modelling a sigmoid decrease instead of increase, so the only different is that the slope parameter has a different sign. The relationship (Yadav, Gopalacharyulu, et al., 2015) between my drug response parameters is:

$$x = m \left(\frac{y - E^{min}}{E^{max} - y} \right)^{-\lambda}$$

Given enough data points we can calculate the best fit for our parameters in the above formula. I make an additional assumption here, which is that a drug that is infinitely diluted will show no inhibition of cell growth (or full viability) by setting the parameter E^{max} to 1 instead of optimising it between 0 and 1.

Using two drugs and denoting the combination treatment with $x_i$and the corresponding single agent treatment with $\tilde{x}_i$, we get the combination index CI that is 1 under Loewe's additivity assumption, smaller than 1 for synergistic combinations, and larger than 1 for antagonistic combinations:

$$\frac{x_1}{\tilde{x}_1} + \frac{x_2}{\tilde{x}_2} = CI$$

If two drugs act in an additive manner, the combination index CI is 1. If it is smaller than 1 the drugs act synergistically, or if it is larger than 1 antagonistically. To calculate our expected response y_{Loewe} given treatment with two drugs at concentration x_1 and x_2, we set the combination index to 1 and solve for y_{Loewe} numerically (Yadav, Wennerberg, et al., 2015):

$$1 - \frac{x_1}{m_1 \left(\frac{y_{Loewe} - E_1^{min}}{E_1^{max} - y_{Loewe}} \right)^{-\lambda_1}} - \frac{x_2}{m_2 \left(\frac{y_{Loewe} - E_2^{min}}{E_2^{max} - y_{Loewe}} \right)^{-\lambda_2}} = 0$$

The way I do this is to use the bisect function of SciPy between 0 and $1 - 10^{-5}$, where the the returned value of the function is smaller than 0 for the input close to 1 and greater than 0 for the input 0. Having those two different signs, the solver divides the interval between the two input values iteratively until it reaches a value for y_{Loewe}that makes the function approximate 0 up to a default tolerance.

I calculate the expected viability y_{Loewe} for the whole matrix of same drug concentration as was used in the combination screening. This way, I get a drug response surface that corresponds to a drug interaction that is neither synergistic nor antagonistic. The volume between the calculated and measured surface then corresponds to an observed interaction effect: If we see fewer cells survive than we would expect

under the assumption of an additive effect we found a synergistic combination, or an antagonistic one if vice versa.

$$V = \int_{x_2^{min}}^{x_2^{max}} \int_{x_1^{min}}^{x_1^{max}} E^{expected} - E^{observed} dx_1 dx_2$$

In practice, I integrate the volume below each of the surfaces separately using the 2D trapezoidal rule by first defining the synergy score s at any given point as

$$s_{m,n} = E^{expected} - E^{observed}$$

Then calculating the volume

$$V = \frac{1}{4} \sum_{j \in x_2} \sum_{i \in x_1} s_{i,j} + s_{i+1,j} + s_{i,j+1} + s_{i+1,j+1} \Delta x_1 \Delta x_2$$

Subtract the measured from the calculated volume, and normalize the volume I obtain by the area of the matrices

$$S = \sum_{x_1} \sum_{x_2} \Delta x_1 \Delta x_2$$

And (optionally) obtain the overall synergy score S by dividing the volume by the area

$$S = \frac{V}{A}.$$

5.3.3 *Single-agent drug response curves*

Starting from the combinations listed in table 11, our collaborators in Hamburg agreed to screen their efficacy in the MFC-7 cell line. This first requires single-agents measurements to be able to fit the Hill parameters for each drug in this cell line, as we would expect them not to be identical with the GDSC due to a different experimental setup. This was done in 384 well plate, seeding cells and letting them attach for 24 hours, then treating them with a drug for 72 (if it was the treatment drug) or 96 hours (if it was the sensitiser), and measuring the number of surviving cells compared to untreated wells using Promega CellTiter Glo as a readout.

Starting from the raw measurement intensity, I calculated the fraction of surviving cells (blue dots) in each of the treated wells for each of the drugs for both 72 (49) and 96 hour treatment (figure 50). I used the Hill equation (first formula in section 5.3.2 and blue line in figures) to fit a dose-response curve to each drug, with the condition that the maximum viability (E_{max} in the above formulas) is 1, *i.e.* the

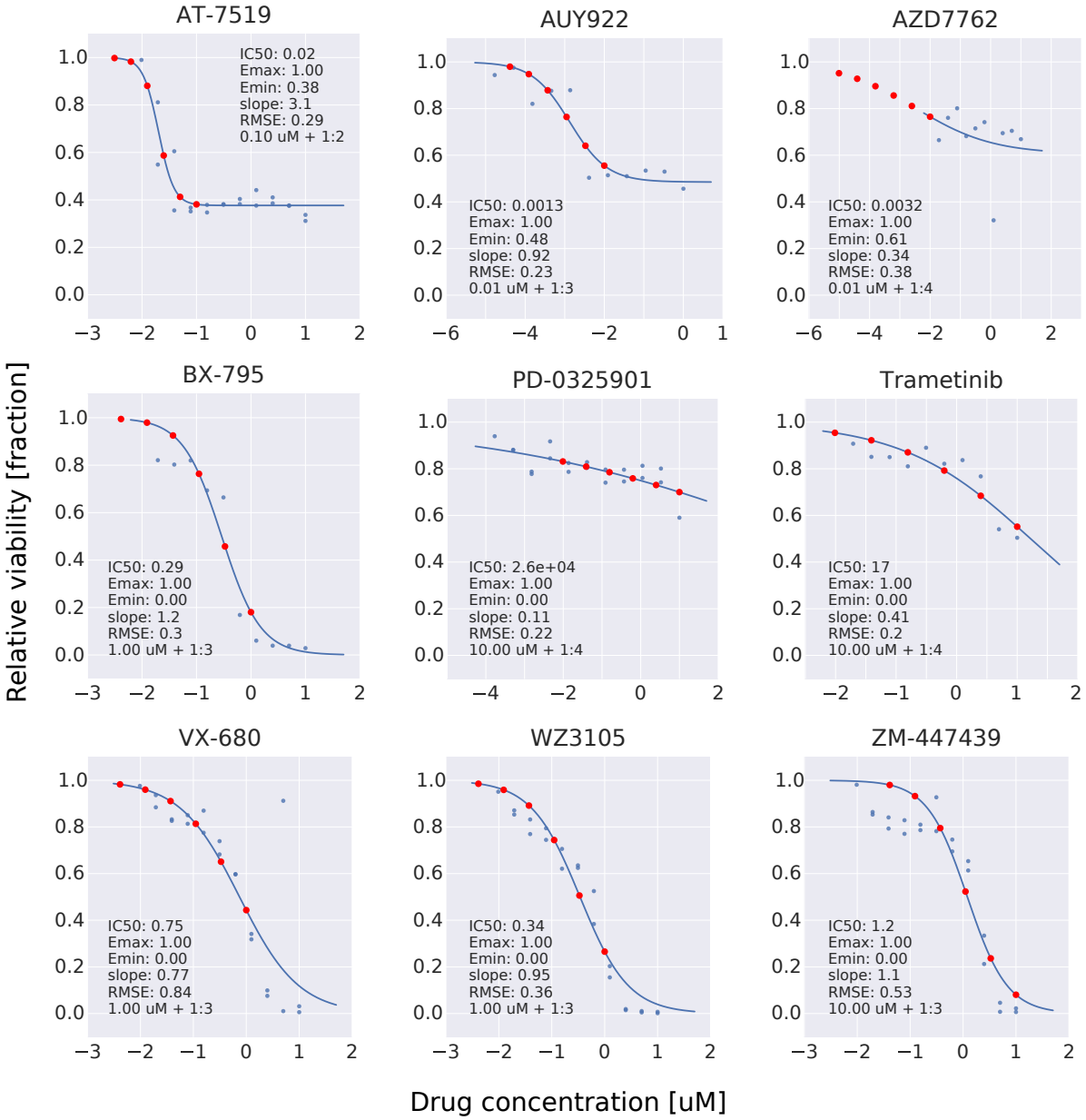

Figure 49: Screening results and fit of drug-response curves for 72 hours of treatment. Blue dots are measured fraction of surviving cells at a given concentration with the blue line as fit. Red dots are the suggested dilutions for this drug in a combination. IC_{50}, E_{min}, E_{max}, steepness of slope and root mean square error (RMSE) of fit as indicated. Starting concentration and dilutions of suggested concentration for combination as indicated.

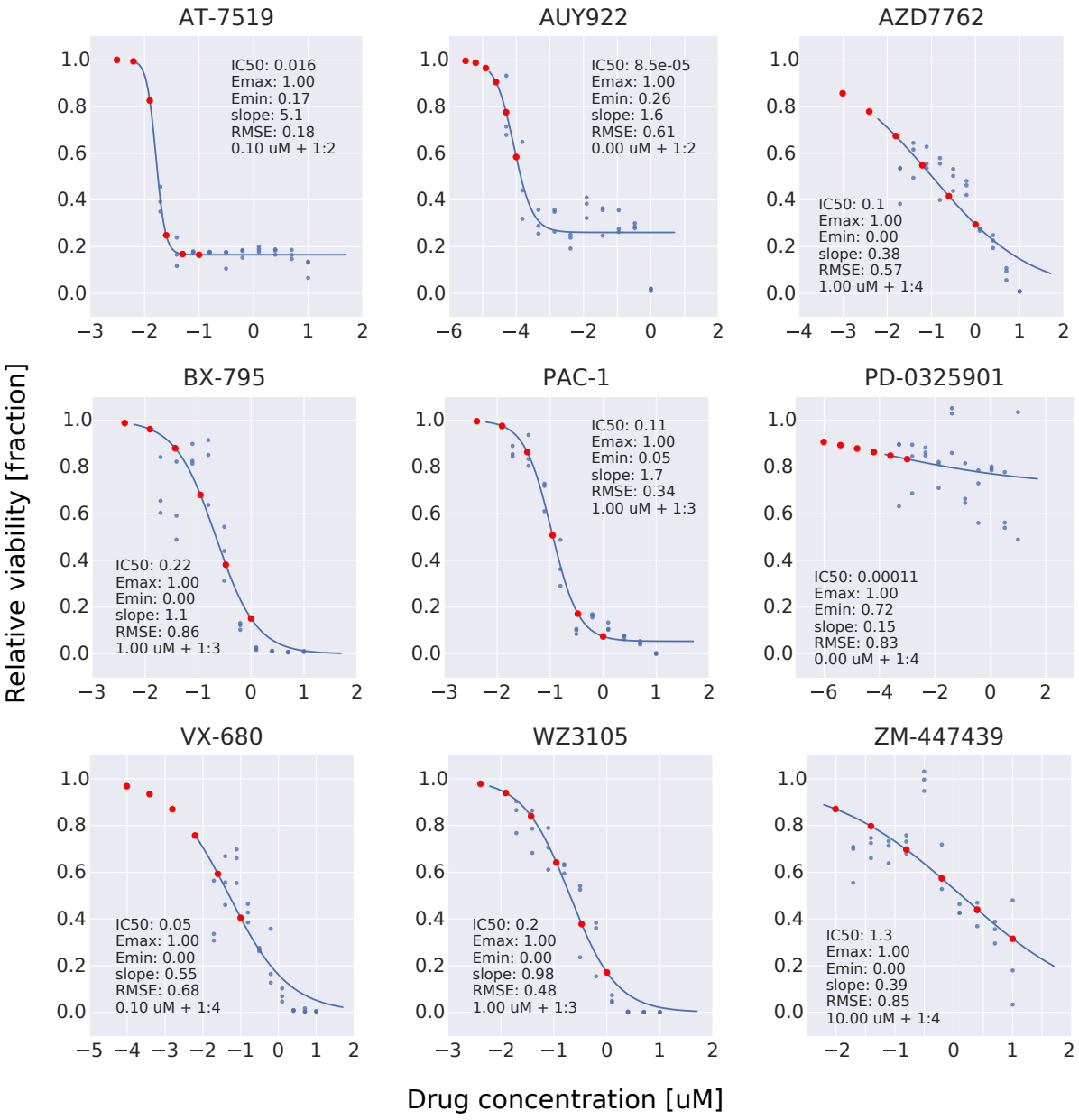

Figure 50: Screening results and fit of drug-response curves for 96 hours of treatment. Legend as in figure 49.

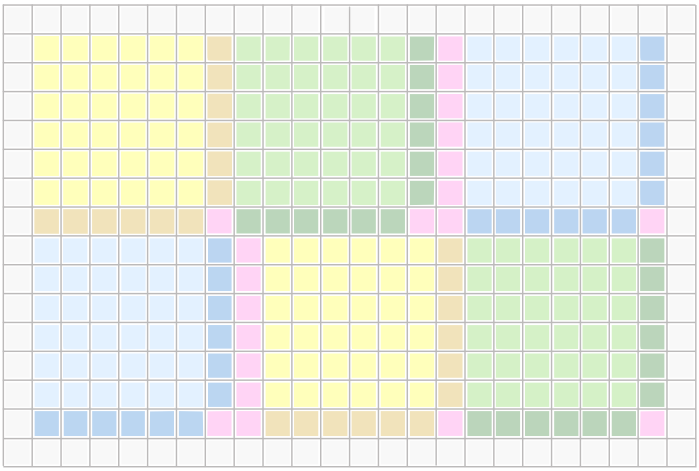

Figure 51: Suggested layout of a 384 well plate to screen six different combinations and have a single-agent control for each on the same plate. Red, yellow, and blue squares correspond to combination dilutions, with darker shades indicating single-agent dilutions as control. Pink indicates growth control without treatment. No measurements on wells on edges and corners.

drug should not have an effect on the viability of the cell given in an infinite dilution. Based on this, I suggested dilutions to use in the combination screening (red dots in figures). Those were designed to start at multiples of 10 uM above the IC_{50} and have a total of 6 dilutions reach below the IC_{10}, but not force stronger dilutions and 1 in 4. The combinations should then be screened in $6x6$ dilutions with a total of 4 combinations including single-agent control on the same plate (figure 51).

5.3.4 *Synergy of drug combinations*

Our collaborators screened 2D dilutions of two different drugs on a $11x11$ grid starting at a concentration of 1 or 10 uM and dilution 2-fold or 3-fold. This takes up more space on the plate than our proposed layout, but it should still capture large parts of the single agent drug response curves for both drugs for all the suggested combinations in table 11. The protocol was to seed cells and let them attach for 24 hours, then treating with the sensitiser, then treating with the treatment drug after another 24 hours, and after 72 more hours measuring the number of surviving cells compared to untreated wells using Promega CellTiter Glo as a readout.

In figures 52-54 (and appendix section C.3), I show representative examples of the results of the screening experiment (left) vs. the calculated viability if the drugs did not interact (middle) and their difference (right). There are multiple aspects to these plots. If the screening con-

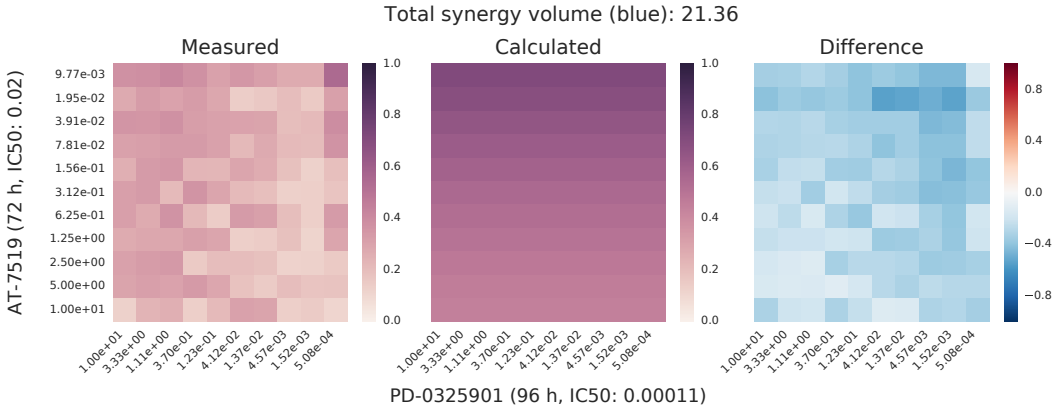

Figure 52: Combination of PD-0325901 as sensitiser and AT-7519 as treatment drug. Both axes represent drug concentrations in micro-molar.

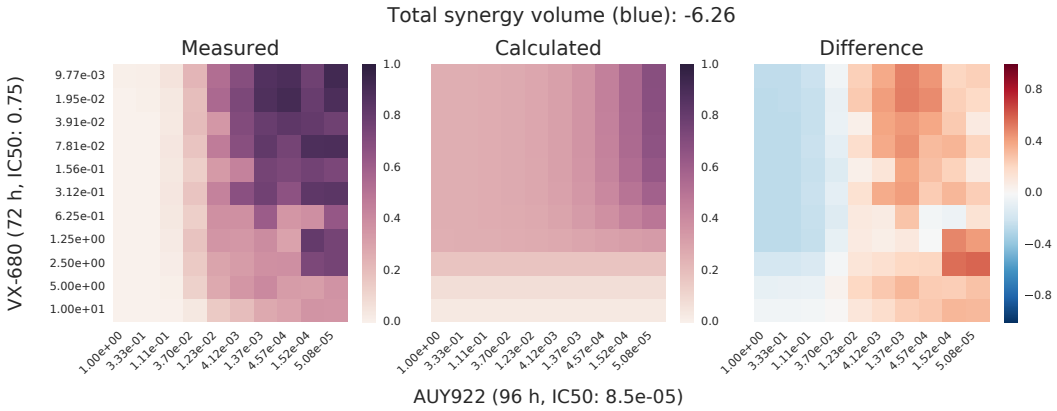

Figure 53: Combination of AUY922 as sensitiser and VX-680 as treatment drug. Both axes represent drug concentrations in micro-molar.

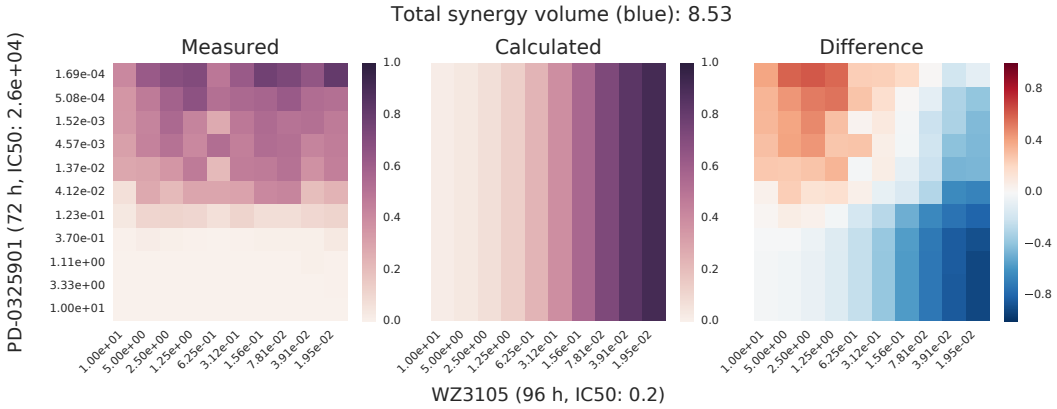

Figure 54: Combination of WZ3105 as sensitiser and PD-0325901 as treatment drug. Both axes represent drug concentrations in micro-molar.

centration using the extended grid did indeed capture the dose-response curve of the single-agent dilutions, the middle panel with the calculated combination efficacy under the null hypothesis (Loewe additivity–no interaction, efficacy of both drugs combined is assuming one drug is a dilution of the other) should show the whole range viability, from bright yellow (all cells are killed) to dark violet (all cells survive). This is true for some of the combinations (figures 53-54) but not others (figure 52).

Another aspect is that even though we do not have single-agent dilutions as part of this matrix, the dose response curve of one drug with the lowest dilution of the other drug should roughly correspond to the single-agent dose response curve. This again is true in some cases (WZ3105 in figure 54), but not others (AUY922 in figure 53).

Finally, my predictions are about synergistic combinations so I would expect the combination of two drugs at various dilutions to kill more cells than the expected null model of interaction. This will of course not always happen, but it should happen more often than expected by random chance. Also for this difference, I find some examples where this is the case (figure 52) but not others where a drug essentially seems to stop working under a certain concentration that does not fit the single-agent dose response curve (figure 53). Interestingly, for some cases (figure 54), I observe that the combination response is driven by the opposite drug than expected from the single-agent curves, leading to strongly synergistic and antagonistic corners of the dilution. Unfortunately, I can at this point not determine if this is a real effect or an artefact due to experimental error.

5.4 DISCUSSION

5.4.1 *Original Connectivity Map*

Using the original Connectivity Map for predicting a drug that could sensitise cell lines to another drug has a couple of drawbacks: First, most of the compounds used are not cancer specific. They do most likely not target a specific kinase but rather a high-level biological process or set of processes. This is not an issue by itself by our prior belief about the potential in treating cancer is lower than if the drug was already designed to treat a specific aspect or type of cancer. Second, gene expression changes are derived from treatment with a high drug concentration, almost exclusively 10 uM. This is a concentration range where a drug that apart from its primary effect has secondary targets will most likely hit them as well, even if the physiological concentration upon drug treatment would never reach this level. In this case, we would observe gene expression changes that are due to both the primary effect but also an unknown number and strength of secondary

effects. Third, most of the perturbations are done only in MCF-7 or a low number of cell lines. The usual time is either 6 or 24 hours.

What speaks for the original Connectivity Map is that with MANTRA 22,000 genes were measured and the compound-level signatures have been merged using informative distances (that is, weighted by the gene expression distance between several experiments, to for instance not have an overly strong contribution of the MCF-7 signature if there were three experiments using MCF-7 and one in PC-3).

5.4.2 *Pan-cancer view of the LINCS*

The LINCS Connectivity Map provides us with an opportunity to investigate the possible effect of combining two cancer drugs, with a total of 150 drugs that both have drug sensitivity estimates in the GDSC and drug response signatures in the LINCS. On top of that, it enabled me to use the drug response signature in combination with the sensitivity estimate to regress out overlapping expression changes between the two. This would not have been possible with the original Connectivity Map, but it would also not have been possible using Gene Set Enrichment Analysis. My analysis has shown one potential approach to effectively remove the effect of overlapping signatures or pathway cross-talk more generally. The available literature support for those models that I could not find for the naive models are evidence of that.

Of course there are some open questions: These are preliminary results, examples are shown where a quick search turned up additional supporting evidence. There are some parameters in the method that should be optimised (LINCS projection, signature creation cutoff). Right now, I only use GDSC drugs in LINCS. There may be a reason to include others as well.

5.4.3 *Tissue-specific models and validation*

The LINCS Connectivity Map also provides us with an opportunity of analysing gene expression response upon drug treatment in a tissue-specific manner while not limiting the set to a single cell line as I would have needed to do in case of the first Connectivity Map. There is five breast cancer and one non-cancerous breast line in the LINCS, with a total of 19,000 perturbations. However, there is no way of telling whether a common signature, or one of the non-cancerous cell line, and which projection etc. would have been the best choice to predict possible drug combinations. Given this limitation, selecting the most consistently synergistically predicted combinations made sense.

Unfortunately, the actual experimental setup seems a lot more error-prone than anticipated. While I suggested the screening concentration

based on single-agent response curves to maximise the information we could obtain by using a relatively small grid and have simultaneously having the single-agent dilutions as a control, the actual setup that was used for screening separated combinations from single-agent control again. This is not a problem if the assay is stable enough to yield comparable Hill parameters, but this does not seem to be the case: in the combination screening, the likely IC_{50} that I can best estimate along a dilution series of a drug by looking at the minimal concentration of the other does often not fit well to the previously estimated single-agent curves. The only way I see to fix this is to assess stability of the assay and influencing factors of the Hill parameters first in order to minimise the batch effect. We might still get another batch of experimental results that takes this into account. This will, however, be past the submission date of this thesis.

6

CONCLUSIONS

Signalling pathways have long been studied in the context of cancer as well as other diseases. Because direct measurements of those are not widely available, significant effort has been devoted to extracting predictive and reliable biomarkers reflecting their status from gene expression data. This has either been done by mapping the expression level of pathway components, or by defining signature genes that differentiate between two conditions. In this thesis, I have shown the utility and robustness of gene expression signatures for inferring signalling activity as well as potential drug combinations. I started with characterising the GDSC cancer cell line panel using gene sets of Gene Ontology categories and Reactome pathways. I showed that one needs to be cautious when interpreting the top associations with drug response, as the process depicted will often likely not be the process that causes sensitivity or resistance to a given drug. As a way to solve this, I showed the advantage of pre-selecting interesting gene sets that are changing upon pathway stimulation using the SPEED platform (Parikh et al., 2010).

I assembled a comprehensive and robust set of consensus gene expression signatures derived from pathway perturbations, which enabled me to detect pathway-specific footprints of signalling activity. I provided the first large-scale comparison between these signatures and state of the art pathway methods in both patient data and pharmacogenomic drug screenings (The Cancer Genome Atlas and Sanger's Genomics of Drug Sensitivity in Cancer, respectively). I found that consensus signatures of perturbations better recover many well-known driver mutations in terms of their expected impact on pathway activity, provide more associations with drug response than driver mutations, and more clearly distinguish between oncogenic and tumour-suppressive pathways for patient survival. Furthermore, I showed that my signatures can be used in combinations with driver mutations to yield better biomarkers for drug indications than mutations alone.

I showed that the same signatures, computed per drug in the LINCS Connectivity Map can be used to predict synergistic and antagonistic drug combinations using signature matching in conjunction with drug sensitivity data. I developed a novel way of computing those associations while taking into account possible pathway cross-talk between the two drugs. While the experimental validation remains to be con-

firmed, literature evidence strongly supported the validity of this approach.

On the technical side, the signatures I have worked with were largely based z-score coefficients of linear models and not gene sets and GSEA. While GSEA has been applied to virtually every possible biological context and we know very well how it behaves (in terms of distribution of scores, appropriate null models, continuous unimodal scores in GSVA, leading edge analysis, etc.). However, the overall knowledge and possibilities of linear models reach far beyond what is possible with GSEA. Adding a covariate to signature matching is just one example of that. It is my impression that those kinds of models will expand again from ANOVAs, Genome Wide Association Studies (GWAS) and eQTLs (genomic variants with impact on gene expression, or expression quantitative trait loci) more into generally applicable analyses of gene expression, because it enables us to apply many more techniques and tools to solve problems such as confounding variables, batch correction, latent variable extraction, etc. A good example of a tool that performs theses tasks is PEER (Stegle et al., 2012), used to regress out known and unknown covariates for genomic associations, while finding correlated latent variables to explain phenotypes.

On the biological side, I expect the signatures I derived (both for pathways and for individual drugs) to in the future be used as tools to interrogate the functional impact of mutations, as well as inference of signalling activity from gene expression for other purposes. As I addressed the issues of post-translational control that common pathway methods do not take into account as well as the context-specificity of single-condition signatures, I believe that those signatures (again, both for pathways and individual drugs) can in the future be used in the pre-clinical as well as clinical setting either as biomarkers for drug indication or patient survival, but also as tools to interrogate the basic biology that drives those processes.

BIBLIOGRAPHY

Alexandrov, Ludmil B et al. (2013). "Signatures of mutational processes in human cancer". en. In: *Nature* 500.7463, pp. 415–421.

Alizadeh, Ash A et al. (2015). "Toward understanding and exploiting tumor heterogeneity". In: *Nat. Med.* 21.8, pp. 846–853.

Allen, Lee F, Judith Sebolt-Leopold, and Mark B Meyer (2003). "CI-1040 (PD184352), a targeted signal transduction inhibitor of MEK (MAPKK)". In: *Semin. Oncol.* 30.5 Suppl 16, pp. 105–116.

Anders, Simon and Wolfgang Huber (2010). "Differential expression analysis for sequence count data". en. In: *Genome Biol.* 11.10, R106.

Aoe, Keisuke et al. (2006). "Expression of vascular endothelial growth factor in malignant mesothelioma". In: *Anticancer Res.* 26.6C, pp. 4833–4836.

Ashburner, Michael et al. (2000). "Gene Ontology: tool for the unification of biology". In: *Nat. Genet.* 25.1, pp. 25–29.

Auerbach, Raymond K, Bin Chen, and Atul J Butte (2013). "Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool". en. In: *Bioinformatics* 29.15, pp. 1922–1924.

Babur, Özgün et al. (2015). "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations". In: *Genome Biol.* 16, p. 45.

Baggerly, Keith A and Kevin R Coombes (2009). "Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology". In: *Ann. Appl. Stat.* 3.4, pp. 1309–1334.

Bailey, T L and C Elkan (1995). "The value of prior knowledge in discovering motifs with MEME". en. In: *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3, pp. 21–29.

Bailey, Timothy L et al. (2009). "MEME SUITE: tools for motif discovery and searching". en. In: *Nucleic Acids Res.* 37.Web Server issue, W202–8.

Baker, Monya (2015). "Reproducibility crisis: Blame it on the antibodies". In: *Nature* 521.7552, pp. 274–276.

– (2016). "Biotech giant publishes failures to confirm high-profile science". In: *Nature* 530.7589, p. 141.

Bansal, Mukesh et al. (2007). "How to infer gene networks from expression profiles". en. In: *Mol. Syst. Biol.* 3.1.

Barretina, Jordi et al. (2012). "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity". en. In: *Nature* 483.7391, pp. 603–607.

Barrett, Tanya, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar (2007). "NCBI GEO: mining tens of millions of expression profiles—database and tools update". In: *Nucleic Acids Res.* 35.suppl 1, pp. D760–D765.

Barrett, Tanya, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, Kimberly A Marshall, et al. (2009). "NCBI GEO: archive for high-throughput functional genomic data". en. In: *Nucleic Acids Res.* 37.Database issue, pp. D885–90.

Basso, Katia et al. (2005). "Reverse engineering of regulatory networks in human B cells". en. In: *Nat. Genet.* 37.4, pp. 382–390.

Begley, C Glenn and Lee M Ellis (2012). "Drug development: Raise standards for preclinical cancer research". en. In: *Nature* 483.7391, pp. 531–533.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *J. R. Stat. Soc. Series B Stat. Methodol.* 57.1, pp. 289–300.

Ben-Kiki, Oren, Clark Evans, and Brian Ingerson (2005). "YAML Ain't Markup Language (YAML™) Version 1.1". In: *yaml. org, Tech. Rep.*

Bensimon, Ariel, Albert J R Heck, and Ruedi Aebersold (2012). "Mass Spectrometry–Based Proteomics and Network Biology". In: *Annu. Rev. Biochem.* 81.1, pp. 379–405.

Bild, Andrea H, Anil Potti, and Joseph R Nevins (2006). "Linking oncogenic pathways with therapeutic opportunities". en. In: *Nat. Rev. Cancer* 6.9, pp. 735–741.

Bild, Andrea H, Guang Yao, et al. (2005). "Oncogenic pathway signatures in human cancers as a guide to targeted therapies". en. In: *Nature* 439.7074, pp. 353–357.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3, pp. 993–1022.

Bradbury, Andrew and Andreas Plückthun (2015). "Reproducibility: Standardize antibodies used in research". In: *Nature* 518.7537, pp. 27–29.

Bray, Nicolas L et al. (2016). "Near-optimal probabilistic RNA-seq quantification". In: *Nat. Biotechnol.* 34.5, pp. 525–527.

Brazma, Alvis et al. (2001). "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data". en. In: *Nat. Genet.* 29.4, pp. 365–371.

Brown, P O and D Botstein (1999). “Exploring the new world of the genome with DNA microarrays”. en. In: *Nat. Genet.* 21.1 Suppl, pp. 33–37.

Burnett, G and E P Kennedy (1954). “The enzymatic phosphorylation of proteins”. en. In: *J. Biol. Chem.* 211.2, pp. 969–980.

Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium (2015). “Pharmacogenomic agreement between two cancer cell line data sets”. en. In: *Nature* 528.7580, pp. 84–87.

Carvalho, Benilton S and Rafael A Irizarry (2010). “A framework for oligonucleotide microarray preprocessing”. In: *Bioinformatics* 26.19, pp. 2363–2367.

Chapman, Paul B et al. (2011). “Improved survival with vemurafenib in melanoma with BRAF V600E mutation”. en. In: *N. Engl. J. Med.* 364.26, pp. 2507–2516.

Chen, Edward Y et al. (2013). “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool”. In: *BMC Bioinformatics* 14, p. 128.

Chen, Li, George Wu, and Hongkai Ji (2011). “hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data”. en. In: *Bioinformatics* 27.10, pp. 1447–1448.

Cheung, Hiu Wing et al. (2011). “Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer”. en. In: *Proceedings of the National Academy of Sciences* 108.30, pp. 12372–12377.

Chibon, Frederic (2013). “Cancer gene expression signatures – The rise and fall?” In: *Eur. J. Cancer* 49.8, pp. 2000–2009.

Ciriello, Giovanni et al. (2012). “Mutual exclusivity analysis identifies oncogenic network modules”. en. In: *Genome Res.* 22.2, pp. 398–406.

Ciuffreda, Ludovica et al. (2009). “Growth-inhibitory and antiangiogenic activity of the MEK inhibitor PD0325901 in malignant melanoma with or without BRAF mutations”. In: *Neoplasia* 11.8, pp. 720–731.

Crick, F H (1958). “On protein synthesis”. en. In: *Symp. Soc. Exp. Biol.* 12, pp. 138–163.

Croce, Carlo M (2008). “Oncogenes and cancer”. en. In: *N. Engl. J. Med.* 358.5, pp. 502–511.

Croft, David et al. (2011). “Reactome: a database of reactions, pathways and biological processes”. In: *Nucleic Acids Res.* 39.Database issue, pp. D691–7.

Dabbish, Laura et al. (2012). “Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository”. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW ’12. New York, NY, USA: ACM, pp. 1277–1286.

Danial, N N and P Rothman (2000). “JAK-STAT signaling activated by Abl oncogenes”. In: *Oncogene* 19.21, pp. 2523–2531.

Debouck, C and P N Goodfellow (1999). “DNA microarrays in drug discovery and development”. en. In: *Nat. Genet.* 21.1 Suppl, pp. 48–50.

Di Leonardo, A et al. (1994). “DNA damage triggers a prolonged p53-dependent G1 arrest and long-term induction of Cip1 in normal human fibroblasts”. en. In: *Genes Dev.* 8.21, pp. 2540–2551.

Domon, Bruno and Ruedi Aebersold (2006). “Mass spectrometry and protein analysis”. en. In: *Science* 312.5771, pp. 212–217.

Drier, Yotam, Michal Sheffer, and Eytan Domany (2013). “Pathway-based personalized analysis of cancer”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 110.16, pp. 6388–6393.

Drukala, Justyna et al. (2010). “ROS accumulation and IGF-IR inhibition contribute to fenofibrate/PPARalpha -mediated inhibition of glioma cell motility in vitro”. In: *Mol. Cancer* 9, p. 159.

Du, Pan, Warren A Kibbe, and Simon M Lin (2008). “lumi: a pipeline for processing Illumina microarray”. en. In: *Bioinformatics* 24.13, pp. 1547–1548.

Duggan, D J et al. (1999). “Expression profiling using cDNA microarrays”. en. In: *Nat. Genet.* 21.1 Suppl, pp. 10–14.

Dutta, Pranabananda and Willis X Li (2013). “Role of the JAK-STAT Signalling Pathway in Cancer”. In: *eLs*.

Elzi, David J et al. (2012). “Wnt antagonist SFRP1 functions as a secreted mediator of senescence”. In: *Mol. Cell. Biol.* 32.21, pp. 4388–4399.

Farmer, Hannah et al. (2005). “Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy”. en. In: *Nature* 434.7035, pp. 917–921.

Ferlay, Jacques et al. (2010). “Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008”. en. In: *Int. J. Cancer* 127.12, pp. 2893–2917.

Fields, Stanley and Mark Johnston (2005). “Whither Model Organism Research?” en. In: *Science* 307.5717, pp. 1885–1886.

Fogh, J (2014). *The Nude Mouse in Experimental and Clinical Research*. Elsevier Science.

Francies, Hayley E and Mathew J Garnett (2015). “What role could organoids play in the personalization of cancer treatment?” In: *Pharmacogenomics* 16.14, pp. 1523–1526.

Friedberg, Errol C (2008). “Sydney Brenner”. en. In: *Nat. Rev. Mol. Cell Biol.* 9.1, pp. 8–9.

Gajanin, Vesna et al. (2010). “Significance of vascular endothelial growth factor expression in skin melanoma”. In: *Vojnosanit. Pregl.* 67.9, pp. 747–754.

Gao, Jianjiong et al. (2013). “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”. en. In: *Sci. Signal.* 6.269, p. l1.

Garnett, Mathew J et al. (2012). “Systematic identification of genomic markers of drug sensitivity in cancer cells”. en. In: *Nature* 483.7391, pp. 570–575.

Gatza, Michael L, Hsiu-Ni Kung, et al. (2011). “Analysis of tumor environmental response and oncogenic pathway activation identifies distinct basal and luminal features in HER2-related breast tumor subtypes”. In: *Breast Cancer Res.* 13.3, R62.

Gatza, Michael L, Joseph E Lucas, et al. (2010). “A pathway-based classification of human breast cancer”. In: *Proc. Natl. Acad. Sci. U. S. A.* 107.15, pp. 6994–6999.

Gatza, Michael L, Grace O Silva, et al. (2014). “An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer”. In: *Nat. Genet.*

Gautier, Laurent et al. (2004). “affy–analysis of Affymetrix GeneChip data at the probe level”. In: *Bioinformatics* 20.3, pp. 307–315.

Gene Ontology Consortium (2004). “The Gene Ontology (GO) database and informatics resource”. In: *Nucleic Acids Res.* 32.suppl 1, pp. D258–D261.

Gentleman, Robert C et al. (2004). “Bioconductor: open software development for computational biology and bioinformatics”. en. In: *Genome Biol.* 5.10, R80.

Gey, Goea, W D Coffman, and Mary T Kubicek (1952). “Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium”. In: *Cancer research*. Vol. 12. … CANCER RESEARCH PO BOX …, pp. 264–265.

Gobbi, Andrea et al. (2014). “Fast randomization of large genomic datasets while preserving alteration counts”. en. In: *Bioinformatics* 30.17, pp. i617–23.

Gohlmann, H and W Talloen (2009). *Gene Expression Studies Using Affymetrix Microarrays*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press.

Greco, W R, G Bravo, and J C Parsons (1995). “The search for synergy: a critical review from a response surface perspective”. en. In: *Pharmacol. Rev.* 47.2, pp. 331–385.

Gry, Marcus et al. (2009). “Correlations between RNA and protein expression profiles in 23 human cell lines”. en. In: *BMC Genomics* 10, p. 365.

Gupta, Mamta et al. (2012). “Elevated serum IL-10 levels in diffuse large B-cell lymphoma: a mechanism of aberrant JAK2 activation”. In: *Blood* 119.12, pp. 2844–2853.

Haibe-Kains, Benjamin et al. (2012). "A three-gene model to robustly identify breast cancer molecular subtypes". en. In: *J. Natl. Cancer Inst.* 104.4, pp. 311–325.

Halaas, O et al. (2000). "Lipopolysaccharide induces expression of APO2 ligand/TRAIL in human monocytes and macrophages". en. In: *Scand. J. Immunol.* 51.3, pp. 244–250.

Hanahan, Douglas and Lisa M Coussens (2012). "Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment". In: *Cancer Cell* 21.3, pp. 309–322.

Hanahan, Douglas and Robert A Weinberg (2000). "The Hallmarks of Cancer". In: *Cell* 100.1, pp. 57–70.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney (2013). "GSVA: gene set variation analysis for microarray and RNA-seq data". In: *BMC Bioinformatics* 14.7, p. 7.

Hartigan, J A and M A Wong (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". In: *J. R. Stat. Soc. Ser. C Appl. Stat.* 28.1, pp. 100–108.

Hata, Aaron N et al. (2016). "Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition". In: *Nat. Med.*

Haynes, Winston A et al. (2013). "Differential expression analysis for pathways". In: *PLoS Comput. Biol.* 9.3, e1002967.

Heppner, G H and B E Miller (1983). "Tumor heterogeneity: biological implications and therapeutic consequences". en. In: *Cancer Metastasis Rev.* 2.1, pp. 5–23.

Hotta, Y, M Ito, and H Stern (1966). "Synthesis of DNA during meiosis". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 56.4, pp. 1184–1191.

Hubbard, T et al. (2002). "The Ensembl genome database project". In: *Nucleic Acids Res.* 30.1, pp. 38–41.

Hutchison 3rd, C A et al. (1974). "Maternal inheritance of mammalian mitochondrial DNA". en. In: *Nature* 251.5475, pp. 536–538.

Ihaka, Ross and Robert Gentleman (1996). "R: A Language for Data Analysis and Graphics". In: *J. Comput. Graph. Stat.* 5.3, pp. 299–314.

International Cancer Genome Consortium et al. (2010). "International network of cancer genome projects". en. In: *Nature* 464.7291, pp. 993–998.

Iorio, Francesco, Roberta Bosotti, et al. (2010). "Discovery of drug mode of action and drug repositioning from transcriptional responses". en. In: *Proceedings of the National Academy of Sciences* 107.33, pp. 14621–14626.

Iorio, Francesco, Theo A Knijnenburg, et al. (2016). "A Landscape of Pharmacogenomic Interactions in Cancer". en. In: *Cell*.

Iorio, Francesco, Timothy Rittman, et al. (2013). "Transcriptional data: a new gateway to drug repositioning?" In: *Drug Discov. Today* 18.7–8, pp. 350–357.

Iorio, Francesco, Roberto Tagliaferri, and Diego di Bernardo (2009). "Identifying Network of Drug Mode of Action by Gene Expression Profiling". In: *J. Comput. Biol.* 16.2, pp. 241–251.

Iskar, Murat et al. (2010). "Drug-induced regulation of target expression". en. In: *PLoS Comput. Biol.* 6.9.

Iverson, Cory et al. (2009). "RDEA119/BAY 869766: a potent, selective, allosteric inhibitor of MEK1/2 for the treatment of cancer". In: *Cancer Res.* 69.17, pp. 6839–6847.

Jeannot, Victor et al. (2014). "The PI3K/AKT pathway promotes gefitinib resistance in mutant KRAS lung adenocarcinoma by a deacetylase-dependent mechanism". In: *Int. J. Cancer* 134.11, pp. 2560–2571.

Jemal, Ahmedin et al. (2011). "Global cancer statistics". en. In: *CA Cancer J. Clin.* 61.2, pp. 69–90.

Jerby-Arnon, Livnat et al. (2014). "Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality". In: *Cell* 158.5, pp. 1199–1209.

Jones, David T W et al. (2012). "Dissecting the genomic complexity underlying medulloblastoma". en. In: *Nature* 488.7409, pp. 100–105.

Kahlos, K et al. (1999). "Generation of reactive oxygen species by human mesothelioma cells". In: *Br. J. Cancer* 80.1-2, pp. 25–31.

Kanehisa, M and S Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Res.* 28.1, pp. 27–30.

Kanehisa, Minoru et al. (2009). "KEGG for representation and analysis of molecular networks involving diseases and drugs". en. In: *Nucleic Acids Res.* 38.Database issue, pp. D355–60.

Khosravi-Far, R et al. (1995). "Activation of Rac1, RhoA, and mitogen-activated protein kinases is required for Ras transformation". In: *Mol. Cell. Biol.* 15.11, pp. 6443–6453.

Kilic-Eren, Mehtap, Tulin Boylu, and Vedrana Tabor (2013). "Targeting PI3K/Akt represses Hypoxia inducible factor-1a activation and sensitizes Rhabdomyosarcoma and Ewing's sarcoma cells for apoptosis". In: *Cancer Cell Int.* 13.1, pp. 1–8.

Killock, David (2014). "Skin cancer: BRAF and MEK inhibitors[mdash]good news comes in twos!" In: *Nat. Rev. Clin. Oncol.* 11.12, pp. 683–683.

Kim, Nan-Hyung et al. (2007). "Impaired PI3K/Akt activation-mediated NF-kappaB inactivation under elevated TNF-alpha is more vulnerable to apoptosis in vitiliginous keratinocytes". In: *J. Invest. Dermatol.* 127.11, pp. 2612–2617.

Kjellman, Christian et al. (2000). "Expression of TGF-β isoforms, TGF-β receptors, and SMAD molecules at different stages of human glioma". In: *International Journal of Cancer* 89.3, pp. 251–258.

Kloo, Bernhard et al. (2011). "Critical role of PI3K signaling for NF-κB–dependent survival in a subset of activated B-cell–like diffuse large B-cell lymphoma cells". In: *Proceedings of the National Academy of Sciences* 108.1, pp. 272–277.

Kunkel, T A and K Bebenek (2000). "DNA replication fidelity". en. In: *Annu. Rev. Biochem.* 69, pp. 497–529.

Kuwahara, M et al. (2001). "Transforming growth factor beta production by spontaneous malignant mesothelioma cell lines derived from Fisher 344 rats". In: *Virchows Arch.* 438.5, pp. 492–497.

Kwong, Lawrence N et al. (2012). "Oncogenic NRAS signaling differentially regulates survival and proliferation in melanoma". en. In: *Nat. Med.* 18.10, pp. 1503–1510.

Lachmann, Alexander et al. (2010). "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments". en. In: *Bioinformatics* 26.19, pp. 2438–2444.

Lamb, Justin (2007). "The Connectivity Map: a new tool for biomedical research". en. In: *Nat. Rev. Cancer* 7.1, pp. 54–60.

Lamb, Justin et al. (2006). "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease". In: *Science* 313.5795, pp. 1929–1935.

Landry, Jonathan J M et al. (2013). "The genomic and transcriptomic landscape of a HeLa cell line". In: *G3: Genes| Genomes| Genetics* 3.8, pp. 1213–1224.

Langfelder, Peter and Steve Horvath (2008). "WGCNA: an R package for weighted correlation network analysis". en. In: *BMC Bioinformatics* 9.1, p. 559.

Laska, E M, M Meisner, and C Siegel (1994). "Simple designs and model-free tests for synergy". en. In: *Biometrics* 50.3, pp. 834–841.

Law, Charity W et al. (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". en. In: *Genome Biol.* 15.2, R29.

Lee, Daniel D and H Sebastian Seung (2001). "Algorithms for Non-negative Matrix Factorization". In: *Advances in Neural Information Processing Systems 13*. Ed. by T K Leen, T G Dietterich, and V Tresp. MIT Press, pp. 556–562.

Lee, Hun Ju et al. (2012). "The role of JAK pathway dysregulation in the pathogenesis and treatment of acute myeloid leukemia". In: *Clin. Cancer Res.* 19.2, pp. 327–335.

Leon, S P, R D Folkerth, and P M Black (1996). "Microvessel density is a prognostic indicator for patients with astroglial brain tumors". In: *Cancer* 77.2, pp. 362–372.

Levine, Arnold J and Moshe Oren (2009). "The first 30 years of p53: growing ever more complex". en. In: *Nat. Rev. Cancer* 9.10, pp. 749–758.

Li, Bo and Colin N Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". en. In: *BMC Bioinformatics* 12, p. 323.

Lipinski, Christopher A et al. (2012). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings". In: *Adv. Drug Deliv. Rev.* 64, Supplement, pp. 4–17.

Loeb, L A, C F Springgate, and N Battula (1974). "Errors in DNA replication as a basis of malignant changes". en. In: *Cancer Res.* 34.9, pp. 2311–2321.

Loeliger, J and M McCullough (2012). *Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development*. O'Reilly Media.

Loewe, S (1926). "Effect of combinations: mathematical basis of the problem". In: *Arch. Exp. Pathol. Pharmakol.* 114, pp. 313–326.

Long, Georgina V et al. (2014). "Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma". en. In: *N. Engl. J. Med.* 371.20, pp. 1877–1888.

Love, M, S Anders, and W Huber (2014). "Differential analysis of count data–the DESeq2 package". In: *Genome Biol.*

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12, p. 550.

Lyseng-Williamson, Katherine A and Caroline Fenton (2005). "Docetaxel: a review of its use in metastatic breast cancer". In: *Drugs* 65.17, pp. 2513–2531.

Maaten, Laurens van der (2013). "Barnes-Hut-SNE". In: arXiv: `1301.3342 [cs.LG]`.

Maier, Tobias, Marc Güell, and Luis Serrano (2009). "Correlation of mRNA and protein in complex biological samples". en. In: *FEBS Lett.* 583.24, pp. 3966–3973.

Malone, James et al. (2010). "Modeling sample variables with an Experimental Factor Ontology". en. In: *Bioinformatics* 26.8, pp. 1112–1118.

Margolin, Adam et al. (2006). "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context". en. In: *BMC Bioinformatics* 7.Suppl 1, S7.

Marioni, John C et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". en. In: *Genome Res.* 18.9, pp. 1509–1517.

Martincorena, Iñigo et al. (2015). "Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin". In: *Science* 348.6237, pp. 880–886.

Mathelier, Anthony et al. (2013). "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles". en. In: *Nucleic Acids Res.* 42.Database issue, pp. D142–7.

Matys, V et al. (2003). "TRANSFAC®: transcriptional regulation, from patterns to profiles". en. In: *Nucleic Acids Res.* 31.1, pp. 374–378.

Maxwell, P H et al. (1999). "The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis". en. In: *Nature* 399.6733, pp. 271–275.

McCulloch, Scott D and Thomas A Kunkel (2008). "The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases". en. In: *Cell Res.* 18.1, pp. 148–161.

McDermott, Ultan (2015). "Next-generation sequencing and empowering personalised cancer medicine". In: *Drug Discov. Today.*

McKinney, W (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* O'Reilly Media.

Mitelman, F, F Mertens, and B Johansson (1997). "A breakpoint map of recurrent chromosomal rearrangements in human neoplasia". en. In: *Nat. Genet.* 15 Spec No, pp. 417–474.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". en. In: *Nat. Methods* 5.7, pp. 621–628.

Ng, Andrew Y, Michael I Jordan, Yair Weiss, et al. (2002). "On spectral clustering: Analysis and an algorithm". In: *Adv. Neural Inf. Process. Syst.* 2, pp. 849–856.

Nijman, Sebastian M B (2011). "Synthetic lethality: general principles, utility and detection using genetic screens in human cells". en. In: *FEBS Lett.* 585.1, pp. 1–6.

Nowell and PC (1960). "A minute chromosome in human granulocytic leukemia". In: *Science* 132, pp. 1497–1501.

Nylander, S et al. (2012). "Human target validation of phosphoinositide 3-kinase (PI3K)β: effects on platelets and insulin sensitivity, using AZD6482 a novel PI3Kβ inhibitor". In: *J. Thromb. Haemost.* 10.10, pp. 2127–2136.

Olive, Kenneth P et al. (2004). "Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome". en. In: *Cell* 119.6, pp. 847–860.

P A Baeuerle and T Henkel (1994). "Function and Activation of NF-kappaB in the Immune System". In: *Annu. Rev. Immunol.* 12.1, pp. 141–179.

Pacini, Clare et al. (2013). "DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data". en. In: *Bioinformatics* 29.1, pp. 132–134.

Pan, Jingxuan et al. (2005). "Farnesyltransferase inhibitors induce DNA damage via reactive oxygen species in human cancer cells". In: *Cancer Res.* 65.9, pp. 3671–3681.

Parikh, Jignesh R et al. (2010). "Discovering causal signaling pathways through gene-expression patterns". In: *Nucleic Acids Res.* 38.Web Server issue, W109–17.

Parkinson, Helen et al. (2009). "ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression". In: *Nucleic Acids Res.* 37.suppl 1, pp. D868–D872.

Parkinson, H et al. (2007). "ArrayExpress—a public database of microarray experiments and gene expression profiles". In: *Nucleic Acids Res.* 35.suppl 1, pp. D747–D750.

Patro, Rob, Stephen M Mount, and Carl Kingsford (2014). "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms". In: *Nat. Biotechnol.* 32.5, pp. 462–464.

Peck, David et al. (2006). "A method for high-throughput gene expression signature analysis". en. In: *Genome Biol.* 7.7, R61.

Pezuk, J A et al. (2013). "Polo-like kinase 1 inhibition causes decreased proliferation by cell cycle arrest, leading to cell death in glioblastoma". In: *Cancer Gene Ther.* 20.9, pp. 499–506.

Pikarsky, Eli and Yinon Ben-Neriah (2006). "NF-kappaB inhibition: a double-edged sword in cancer?" In: *Eur. J. Cancer* 42.6, pp. 779–784.

Pilato, C M, B Collins-Sussman, and B W Fitzpatrick (2008). *Version Control with Subversion*. O'Reilly Media.

Plummer, Ruth, Christopher Jones, et al. (2008). "Phase I study of the poly(ADP-ribose) polymerase inhibitor, AG014699, in combination with temozolomide in patients with advanced solid tumors". In: *Clin. Cancer Res.* 14.23, pp. 7917–7923.

Plummer, Ruth, Paul Lorigan, et al. (2013). "A phase II study of the potent PARP inhibitor, Rucaparib (PF-01367338, AG014699), with temozolomide in patients with metastatic melanoma demonstrating evidence of chemopotentiation". In: *Cancer Chemother. Pharmacol.* 71.5, pp. 1191–1199.

Portales-Casamar, Elodie et al. (2009). "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles". In: *Nucleic Acids Res.* 38.Database issue, pp. D105–10.

Rajabi, Parvin et al. (2012). "The role of VEGF in melanoma progression". In: *J. Res. Med. Sci.* 17.6, pp. 534–539.

Ramaswamy, Sridhar (2007). "Rational design of cancer-drug combinations". en. In: *N. Engl. J. Med.* 357.3, pp. 299–300.

Reardon, David A et al. (2008). "Glioblastoma multiforme: an emerging paradigm of anti-VEGF therapy". In: *Expert Opin. Biol. Ther.* 8.4, pp. 541–553.

Reardon, Sara (2016). "A mouse's house may ruin experiments". In: *Nature News* 530.7590, p. 264.

Reya, T et al. (2001). "Stem cells, cancer, and cancer stem cells". en. In: *Nature* 414.6859, pp. 105–111.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". en. In: *Bioinformatics* 26.1, pp. 139–140.

Rothenfusser, Simon et al. (2002). "Plasmacytoid dendritic cells: the key to CpG". In: *Hum. Immunol.* 63.12, pp. 1111–1119.

Rubio-Perez, Carlota et al. (2015). "In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities". In: *Cancer Cell* 27.3, pp. 382–396.

Ruiz-Garcia, Ana et al. (2008). "Pharmacokinetics in drug discovery". en. In: *J. Pharm. Sci.* 97.2, pp. 654–690.

Sachs, Norman and Hans Clevers (2014). "Organoid cultures for the analysis of cancer phenotypes". en. In: *Curr. Opin. Genet. Dev.* 24, pp. 68–73.

Sanger, F, S Nicklen, and A R Coulson (1977). "DNA sequencing with chain-terminating inhibitors". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.12, pp. 5463–5467.

Schultz, B E and S I Chan (2001). "Structures and proton-pumping strategies of mitochondrial respiratory enzymes". en. In: *Annu. Rev. Biophys. Biomol. Struct.* 30, pp. 23–65.

Schuster, Stephan C (2008). "Next-generation sequencing transforms today's biology". en. In: *Nat. Methods* 5.1, pp. 16–18.

Schwab, Matthias, Martin Karrenbach, and Jon Claerbout (2000). "Making Scientific Computations Reproducible". In: *Comput. Sci. Eng.* 2.6, pp. 61–67.

Shalon, D, S J Smith, and P O Brown (1996). "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization". en. In: *Genome Res.* 6.7, pp. 639–645.

Shen, Helen (2014). "Interactive notebooks: Sharing the code". en. In: *Nature* 515.7525, pp. 151–152.

Shutes, Adam et al. (2007). "Specificity and mechanism of action of EHT 1864, a novel small molecule inhibitor of Rac family small GTPases". In: *J. Biol. Chem.* 282.49, pp. 35666–35678.

Skeel, R T and S N Khleif (2011). *Handbook of Cancer Chemotherapy*. A Lippincott Williams & Wilkins Handbook. Lippincott Williams & Wilkins.

Sleijfer, S et al. (1998). "Induction of tumor necrosis factor-alpha as a cause of bleomycin-related toxicity". In: *Cancer* 82.5, pp. 970–974.

Smedley, Damian et al. (2009). "BioMart–biological queries made easy". In: *BMC Genomics* 10, p. 22.

Smit, Linda et al. (2004). "Wnt activates the Tak1/Nemo-like kinase pathway". en. In: *J. Biol. Chem.* 279.17, pp. 17232–17240.

Smyth, G K (2005). "limma: Linear Models for Microarray Data". In: *Bioinformatics and Computational Biology Solutions Using R and*

Bioconductor. Statistics for Biology and Health. Springer New York, pp. 397–420.

Stallman, Richard M and Roland McGrath (1991). "GNU Make - A Program for Directing Recompilation". In: *Free Software Foundation*.

Stallman, Richard M, Roland McGrath, and Paul D Smith (2004). *GNU Make: A Program for Directing Recompilation, for Version 3.81*. Free Software Foundation.

Stegle, Oliver et al. (2012). "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses". en. In: *Nat. Protoc.* 7.3, pp. 500–507.

Subramanian, Aravind et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.43, pp. 15545–15550.

Talmadge, James E et al. (2007). "Murine models to evaluate novel and conventional therapeutic strategies for cancer". en. In: *Am. J. Pathol.* 170.3, pp. 793–804.

Tao, Guo-Zhong et al. (2013). "Wnt/β-catenin signaling protects mouse liver against oxidative stress-induced apoptosis through the inhibition of forkhead transcription factor FoxO3". In: *J. Biol. Chem.* 288.24, pp. 17214–17224.

Tarca, Adi Laurentiu et al. (2008). "A novel signaling pathway impact analysis". en. In: *Bioinformatics* 25.1, pp. 75–82.

– (n.d.). "A Novel Signaling Pathway Impact Analysis (SPIA)". In:

The Cancer Genome Atlas Research Network et al. (2013). "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nat. Genet.* 45.10, pp. 1113–1120.

Thomas, Dave and Andy Hunt (2003). *Pragmatic Version Control Using CVS*. The Pragmatic Programmers.

Tibes, Raoul et al. (2006). "Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells". en. In: *Mol. Cancer Ther.* 5.10, pp. 2512–2521.

Trapnell, Cole, Adam Roberts, et al. (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". en. In: *Nat. Protoc.* 7.3, pp. 562–578.

Trapnell, Cole, Brian A Williams, et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". en. In: *Nat. Biotechnol.* 28.5, pp. 511–515.

Twombly, Renee (2005). "Cancer surpasses heart disease as leading cause of death for all but the very elderly". en. In: *J. Natl. Cancer Inst.* 97.5, pp. 330–331.

Vainchenker, W and S N Constantinescu (2012). "JAK/STAT signaling in hematological malignancies". In: *Oncogene* 32.21, pp. 2601–2613.

Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *J. Mach. Learn. Res.* 9.85, pp. 2579–2605.

Van Rossum, Guido and Fred L Drake Jr (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Vaske, Charles J et al. (2010). "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM". en. In: *Bioinformatics* 26.12, pp. i237–i245.

Vogelstein, Bert and Kenneth W Kinzler (2004). "Cancer genes and the pathways they control". en. In: *Nat. Med.* 10.8, pp. 789–799.

Wadi, Lina et al. (2016). "Impact of knowledge accumulation on pathway enrichment analysis".

Wall, Michael E, Andreas Rechtsteiner, and Luis M Rocha (2003). "Singular Value Decomposition and Principal Component Analysis". en. In: *A Practical Approach to Microarray Data Analysis*. Ed. by Daniel P Berrar, Werner Dubitzky, and Martin Granzow. Springer US, pp. 91–109.

Wang, Kai et al. (2009). "Genome-wide identification of post-translational modulators of transcription factor activity in human B cells". en. In: *Nat. Biotechnol.* 27.9, pp. 829–837.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". en. In: *Nat. Rev. Genet.* 10.1, pp. 57–63.

Warr, Wendy A (2012). "Scientific workflow systems: Pipeline Pilot and KNIME". en. In: *J. Comput. Aided Mol. Des.* 26.7, pp. 801–804.

Watson, J D (1987). "Molecular biology of the gene". In:

Wei, Guo et al. (2006). "Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance". In: *Cancer Cell* 10.4, pp. 331–342.

Weinstein, I Bernard (2002). "Cancer. Addiction to oncogenes–the Achilles heal of cancer". en. In: *Science* 297.5578, pp. 63–64.

Weissmueller, Susann et al. (2014). "Mutant p53 drives pancreatic cancer metastasis through cell-autonomous PDGF receptor β signaling". en. In: *Cell* 157.2, pp. 382–394.

Wickham, Hadley (2011). "ggplot2". In: *WIREs Comp Stat* 3.2, pp. 180–185.

Wickham, Hadley and Romain Francois (2014). "dplyr: A grammar of data manipulation". In: *URL http://CRAN. R-project. org/package= dplyr. R package version 0. 2*.

Wold, Svante, Kim Esbensen, and Paul Geladi (1987). "Principal component analysis". In: *Chemometrics Intellig. Lab. Syst.* 2.1-3, pp. 37–52.

Wolstencroft, Katherine et al. (2013). "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud". en. In: *Nucleic Acids Res.* 41.Web Server issue, W557–61.

Xie, Yihui (2014). "knitr: a comprehensive tool for reproducible research in R". In: *Implement Reprod Res* 1, p. 20.

Xu, Rui and Donald Wunsch 2nd (2005). "Survey of clustering algorithms". en. In: *IEEE Trans. Neural Netw.* 16.3, pp. 645–678.

Yadav, Bhagwan, Peddinti Gopalacharyulu, et al. (2015). "From drug response profiling to target addiction scoring in cancer cell models". In: *Dis. Model. Mech.* 8.10, pp. 1255–1264.

Yadav, Bhagwan, Krister Wennerberg, et al. (2015). "Searching for Drug Synergy in Complex Dose–Response Landscapes Using an Interaction Potency Model". In: *Comput. Struct. Biotechnol. J.* 13, pp. 504–513.

Yamada, Naoshi et al. (1995). "Enhanced expression of transforming growth factor-β and its type-I and type-II receptors in human glioblastoma". In: *International Journal of Cancer* 62.4, pp. 386–392.

Yang, Xiu-Mei et al. (2009). "Role of PI3K/Akt and MEK/ERK in mediating hypoxia-induced expression of HIF-1alpha and VEGF in laser-induced rat choroidal neovascularization". en. In: *Invest. Ophthalmol. Vis. Sci.* 50.4, pp. 1873–1879.

Yap, Timothy A and Paul Workman (2012). "Exploiting the cancer genome: strategies for the discovery and clinical development of targeted molecular therapeutics". In: *Annu. Rev. Pharmacol. Toxicol.* 52.1, pp. 549–573.

Yates, Andrew et al. (2016). "Ensembl 2016". In: *Nucleic Acids Res.* 44.D1, pp. D710–6.

Zecchin, Davide et al. (2013). "BRAF V600E is a determinant of sensitivity to proteasome inhibitors". en. In: *Mol. Cancer Ther.* 12.12, pp. 2950–2961.

Zerbino, Daniel R et al. (2016). "Ensembl regulation resources". In: *Database* 2016.

Zhang, Cen et al. (2013). "Tumour-associated mutant p53 drives the Warburg effect". en. In: *Nat. Commun.* 4, p. 2935.

Zhang, Qisheng et al. (2007). "Small-molecule synergist of the Wnt/beta-catenin signaling pathway". In: *Proc. Natl. Acad. Sci. U. S. A.* 104.18, pp. 7444–7448.

Zhao, Wei et al. (2014). "A New Bliss Independence Model to Analyze Drug Combination Data". en. In: *J. Biomol. Screen.* 19.5, pp. 817–821.

Zhou, Jie et al. (2004). "PI3K/Akt Is Required for Heat Shock Proteins to Protect Hypoxia-inducible Factor 1α from pVHL-independent Degradation". In: *J. Biol. Chem.* 279.14, pp. 13506–13513.

Zhu, Jiajun et al. (2015). “Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth”. In: *Nature* 525.7568, pp. 206–211.

APPENDIX

A ASSOCIATIONS FOR BASELINE METHODS (CHAPTER 2)

A.1 *Drug response with unbiased gene sets*

Associations for mutations

Table A1: Mutations vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
Nutlin-3a	TP53	452	1.426	15.118	2.6e-44	1.8e-39
Dabrafenib	BRAF	66	-3.129	-12.613	8.5e-33	3e-28
PLX4720 (rescreen)	BRAF	75	-1.648	-10.349	2e-23	4.6e-19
PLX4720	BRAF	66	-1.667	-8.801	1.3e-17	2.2e-13
SB590885	BRAF	62	-1.374	-7.811	2.4e-14	3.4e-10
SGC0946	ADCY1	1	-3.357	-6.285	5.8e-10	6.8e-06
VNLG/124	TET2	15	-1.189	-5.996	3.3e-09	3e-05
GSK690693	PTEN	78	-1.125	-5.985	3.5e-09	3e-05
AV-951	FIP1L1	2	-2.581	-5.913	5.3e-09	4e-05
PD-0325901	NRAS	45	-1.436	-5.908	5.7e-09	4e-05
RDEA119 (rescreen)	KRAS	96	-1.062	-5.876	6.7e-09	4.2e-05
RDEA119	NRAS	45	-1.386	-5.761	1.3e-08	7.6e-05
ABT-869	U2AF1	5	-2.116	-5.734	1.5e-08	7.8e-05
VX-11e	BRAF	72	-1.207	-5.664	2.2e-08	0.00011
(5Z)-7-Oxozeaenol	BRAF	73	-1.105	-5.558	3.9e-08	0.00018
AR-42	TBX3	3	4.719	5.518	4.9e-08	0.00021
IOX2	FIP1L1	2	-2.755	-5.470	6.3e-08	0.00026
VX-702	NTN4	1	-5.037	-5.464	6.7e-08	0.00026
PD-173074	RUNX1	2	-3.873	-5.427	8.2e-08	0.0003
AC220	U2AF1	5	-2.117	-5.385	9.9e-08	0.00034
LAQ824	PIK3R1	14	1.636	5.275	1.8e-07	0.00059
NVP-BHG712	BRAF	72	-1.156	-5.266	1.9e-07	0.00059
Gefitinib	EGFR	10	-1.703	-5.206	2.6e-07	0.00079
PD-0325901	KRAS	86	-1.081	-5.116	4.2e-07	0.0012
ABT-888	MEF2C	1	-3.418	-5.086	4.8e-07	0.0013
A-770041	DICER1	1	-8.384	-5.060	8e-07	0.0021
A-770041	APAF1	1	-8.384	-5.060	8e-07	0.0021
T0901317	MET	8	-1.442	-4.937	1e-06	0.0025
LAQ824	FIP1L1	1	5.457	4.875	1.4e-06	0.0033
OSI-930	FIP1L1	2	-2.938	-4.815	1.8e-06	0.0042
WH-4-023	DICER1	1	-8.982	-4.843	2.2e-06	0.0048
WH-4-023	APAF1	1	-8.982	-4.843	2.2e-06	0.0048
FR-180204	BRAF	72	-0.619	-4.760	2.4e-06	0.0048
Cisplatin	PRPF8	1	-5.287	-4.763	2.4e-06	0.0048
AZD-0530	DICER1	1	-5.454	-4.774	3e-06	0.0058
AZD-0530	APAF1	1	-5.454	-4.774	3e-06	0.0058
PD-0332991	RB1	61	1.093	4.680	3.5e-06	0.0067
ABT-869	MYCN	1	-3.842	-4.666	3.7e-06	0.0068
Obatoclax Mesylate	DLG1	4	3.924	4.627	4.5e-06	0.008
PLX4720	NRAS	46	0.875	4.566	6e-06	0.01
Trametinib	NRAS	46	-1.474	-4.564	6e-06	0.01
(5Z)-7-Oxozeaenol	TP53	491	0.491	4.537	6.7e-06	0.011
SGC0946	GNAS	3	-1.412	-4.481	8.7e-06	0.014
RDEA119	KRAS	85	-0.938	-4.454	1e-05	0.015
PXD101, Belinostat	TBX3	3	4.368	4.452	1e-05	0.015
Cetuximab	TJP2	4	-2.280	-4.443	1e-05	0.016
AC220	EIF4A2	1	-3.827	-4.425	1.1e-05	0.016
FTI-277	ARID2	16	-0.874	-4.436	1.1e-05	0.016
XL-880	NRAS	46	0.911	4.418	1.2e-05	0.016
Dasatinib	DICER1	1	-8.850	-4.442	1.3e-05	0.018
A-770041	SETDB1	2	-5.367	-4.438	1.3e-05	0.018
Bleomycin (50 uM)	SMARCA4	40	-1.316	-4.398	1.3e-05	0.018
Dasatinib	APAF1	1	-8.850	-4.442	1.3e-05	0.018

KIN001-260	WNK1	9	-1.422	-4.373	1.4e-05	0.018
Lapatinib	MET	1	-4.538	-4.417	1.5e-05	0.018
Dabrafenib	TP53	467	0.618	4.366	1.5e-05	0.018
ATRA	CDKN1B	2	-3.470	-4.362	1.5e-05	0.018
Lapatinib	PLCG1	1	-4.538	-4.417	1.5e-05	0.018
Nutlin-3a	CTNNB1	16	-1.344	-4.348	1.6e-05	0.019
YM155	HGF	3	5.891	4.331	1.7e-05	0.02
WH-4-023	MAP3K1	2	-5.733	-4.353	1.9e-05	0.021
XL-184	U2AF1	5	-2.336	-4.305	1.9e-05	0.021
GW-2580	ADCY1	1	-2.953	-4.303	1.9e-05	0.021
CP724714	MET	8	-1.472	-4.274	2.2e-05	0.024
Tamoxifen	FIP1L1	2	-2.129	-4.272	2.2e-05	0.024
KIN001-270	WNK1	9	-1.093	-4.267	2.3e-05	0.024
CI-1040	KRAS	90	-0.777	-4.252	2.4e-05	0.025
GSK690693	PIK3CA	76	-0.799	-4.229	2.7e-05	0.027
PLX4720 (rescreen)	NRAS	48	0.704	4.209	2.9e-05	0.029
Dabrafenib	STK4	1	-6.559	-4.166	3.5e-05	0.035
A-770041	MAP3K1	2	-4.973	-4.198	3.7e-05	0.036
SGC0946	SYK	2	-1.624	-4.153	3.7e-05	0.036
FH535	IRF2	1	4.989	4.124	4.2e-05	0.04
TL-1-85	NRAS	47	0.833	4.113	4.4e-05	0.041
NG-25	NRAS	47	0.907	4.098	4.7e-05	0.043
Dasatinib	SETDB1	2	-5.988	-4.139	4.7e-05	0.043
Vorinostat	FKBP5	1	4.078	4.100	4.7e-05	0.043
Obatoclax Mesylate	IRF2	1	6.640	4.091	4.8e-05	0.043
UNC1215	STK4	1	-1.955	-4.080	5e-05	0.044
Trametinib	KRAS	92	-1.147	-4.075	5.1e-05	0.044
GSK690693	MLLT4	16	-1.620	-4.074	5.1e-05	0.044

Associations for Gene Ontology

Table A2: Gene Ontology vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
Trametinib	GO:0030056 hemidesmosome	713	-1.952	-8.504	1.2e-16	1.9e-10
Nutlin-3a	GO:0003283 atrial septum development	676	-1.330	-8.336	4.7e-16	3.1e-10
RDEA119 (rescreen)	GO:0030056 hemidesmosome	698	-1.318	-8.299	5.8e-16	3.1e-10
Nutlin-3a	GO:0071494 cellular response to UV-C	676	-0.976	-7.967	7.4e-15	3e-09
Nutlin-3a	GO:0008625 extrinsic apoptotic signal...	676	-2.038	-7.936	9.2e-15	3e-09
Trametinib	GO:0001948 glycoprotein binding	713	-3.833	-7.819	2e-14	5.4e-09
NPK76-II-72-1	GO:0003723 RNA binding	733	-2.534	-7.757	3.1e-14	7.1e-09
WZ3105	GO:0042908 xenobiotic transport	734	1.223	7.611	8.8e-14	1.8e-08
Nutlin-3a	GO:0006927 transformed cell apoptotic...	676	-1.018	-7.543	1.6e-13	2.8e-08
Trametinib	GO:0002020 protease binding	713	-3.570	-7.450	2.8e-13	4.2e-08
CCT018159	GO:0006144 purine nucleobase metaboli...	715	-1.378	-7.439	3e-13	4.2e-08
Nutlin-3a	GO:0003181 atrioventricular valve mor...	676	-1.272	-7.443	3.2e-13	4.2e-08
GSK-650394	GO:0002181 cytoplasmic translation	680	-1.290	-7.426	3.5e-13	4.4e-08
Trametinib	GO:0032587 ruffle membrane	713	-4.024	-7.396	4.1e-13	4.7e-08
CCT018159	GO:0009168 purine ribonucleoside mono...	715	-0.954	-7.359	5.3e-13	5.7e-08
DMOG	GO:0035879 plasma membrane lactate tr...	692	-1.317	-7.287	9.1e-13	9.1e-08
NPK76-II-72-1	GO:0030529 ribonucleoprotein complex	733	-1.975	-7.266	9.9e-13	9.4e-08
Bleomycin (50 uM)	GO:0048870 cell motility	746	-2.329	-7.254	1.1e-12	9.4e-08
Trametinib	GO:0000188 inactivation of MAPK activity	713	-2.785	-7.227	1.3e-12	1e-07
EHT 1864	GO:0005759 mitochondrial matrix	737	-1.319	-7.227	1.3e-12	1e-07
GSK-650394	GO:0006413 translational initiation	680	-1.704	-7.239	1.3e-12	1e-07
Trametinib	GO:0048870 cell motility	713	-2.564	-7.205	1.5e-12	1.1e-07
NPK76-II-72-1	GO:0005732 small nucleolar ribonucleo...	733	-1.023	-7.204	1.5e-12	1.1e-07
EHT 1864	GO:0042645 mitochondrial nucleoid	737	-0.795	-7.174	1.8e-12	1.2e-07
RDEA119	GO:0030056 hemidesmosome	668	-1.328	-7.188	1.8e-12	1.2e-07
Trametinib	GO:0005925 focal adhesion	713	-3.643	-7.103	3.1e-12	1.9e-07
RDEA119 (rescreen)	GO:0001948 glycoprotein binding	698	-2.398	-7.081	3.6e-12	2.1e-07
RDEA119 (rescreen)	GO:0002020 protease binding	698	-2.349	-7.080	3.7e-12	2.1e-07
Afatinib (rescreen)	GO:0070830 bicellular tight junction ...	727	-2.246	-7.061	4e-12	2.2e-07
NPK76-II-72-1	GO:0006396 RNA processing	733	-1.853	-7.060	4e-12	2.2e-07
Afatinib	GO:0086073 bundle of His cell-Purkinj...	675	-1.199	-7.053	4.5e-12	2.3e-07
Afatinib	GO:0086083 cell adhesive protein bind...	675	-1.199	-7.053	4.5e-12	2.3e-07
Trametinib	GO:0031581 hemidesmosome assembly	713	-1.851	-7.021	5.3e-12	2.6e-07
YK 4-279	GO:0001731 formation of translation p...	627	-1.108	-7.028	5.7e-12	2.7e-07
Afatinib	GO:0070830 bicellular tight junction ...	675	-2.095	-6.988	7e-12	3.2e-07
Trametinib	GO:0051045 negative regulation of mem...	713	-1.739	-6.977	7.1e-12	3.2e-07
NPK76-II-72-1	GO:0071013 catalytic step 2 spliceosome	733	-1.421	-6.949	8.4e-12	3.6e-07
YK 4-279	GO:0005852 eukaryotic translation ini...	627	-1.026	-6.964	8.8e-12	3.7e-07
GSK-650394	GO:0019083 viral transcription	680	-1.521	-6.895	1.3e-11	5.3e-07

YK 4-279	GO:0006446 regulation of translationa...	627	-1.343	-6.893	1.4e-11	5.6e-07
Bleomycin (50 uM)	GO:0016032 viral process	746	-3.292	-6.860	1.5e-11	5.8e-07
RDEA119 (rescreen)	GO:0048870 cell motility	698	-1.683	-6.858	1.6e-11	6.1e-07
Trametinib	GO:0033627 cell adhesion mediated by ...	713	-1.791	-6.831	1.9e-11	7e-07
Gemcitabine	GO:0003723 RNA binding	685	-4.326	-6.823	2e-11	7.3e-07
RDEA119 (rescreen)	GO:0005925 focal adhesion	698	-2.434	-6.824	2e-11	7.3e-07
WZ3105	GO:0006855 drug transmembrane transport	734	1.694	6.811	2.1e-11	7.3e-07
Etoposide	GO:0003723 RNA binding	696	-3.125	-6.806	2.2e-11	7.5e-07
Etoposide	GO:0019083 viral transcription	696	-1.765	-6.812	2.2e-11	7.4e-07
PF-562271	GO:0006413 translational initiation	680	-1.040	-6.802	2.3e-11	7.6e-07
Trametinib	GO:0071438 invadopodium membrane	713	-1.554	-6.790	2.4e-11	7.7e-07
YM155	GO:0006805 xenobiotic metabolic process	706	4.456	6.796	2.4e-11	7.6e-07
Gemcitabine	GO:0006189 'de novo' IMP biosynthetic...	685	-1.517	-6.789	2.5e-11	7.8e-07
GSK-650394	GO:0005840 ribosome	680	-1.754	-6.757	3.1e-11	9.2e-07
T0901317	GO:0005852 eukaryotic translation ini...	728	-0.596	-6.749	3.1e-11	9.2e-07
RDEA119 (rescreen)	GO:0042127 regulation of cell prolife...	698	-2.920	-6.757	3.1e-11	9.2e-07
PF-562271	GO:0006446 regulation of translationa...	680	-1.039	-6.745	3.4e-11	9.6e-07
YK 4-279	GO:0016282 eukaryotic 43S preinitiati...	627	-0.980	-6.755	3.4e-11	9.6e-07
RDEA119 (rescreen)	GO:0097296 activation of cysteine-typ...	698	-1.254	-6.719	3.9e-11	1.1e-06
Thapsigargin	GO:0001825 blastocyst formation	680	-1.743	-6.718	4e-11	1.1e-06
Etoposide	GO:0019058 viral life cycle	696	-2.239	-6.707	4.2e-11	1.1e-06
T0901317	GO:0000027 ribosomal large subunit as...	728	-0.622	-6.689	4.6e-11	1.2e-06
NPK76-II-72-1	GO:0000166 nucleotide binding	733	-2.512	-6.677	5e-11	1.3e-06
Afatinib	GO:0005923 bicellular tight junction	675	-2.865	-6.684	5e-11	1.3e-06
PF-562271	GO:0033290 eukaryotic 48S preinitiati...	680	-0.759	-6.684	5e-11	1.3e-06
PF-562271	GO:0006412 translation	680	-1.227	-6.682	5.1e-11	1.3e-06
EHT 1864	GO:0002161 aminoacyl-tRNA editing act...	737	-0.591	-6.662	5.4e-11	1.3e-06
GSK-650394	GO:0000184 nuclear-transcribed mRNA c...	680	-1.578	-6.670	5.5e-11	1.3e-06
CCT018159	GO:0006164 purine nucleotide biosynth...	715	-0.891	-6.655	5.8e-11	1.4e-06
Vorinostat	GO:0003014 renal system process	677	0.934	6.647	6.4e-11	1.5e-06
T0901317	GO:0033290 eukaryotic 48S preinitiati...	728	-0.569	-6.636	6.5e-11	1.5e-06
Epothilone B	GO:0000463 maturation of LSU-rRNA fro...	689	-1.042	-6.638	6.7e-11	1.5e-06
YK 4-279	GO:0033290 eukaryotic 48S preinitiati...	627	-0.950	-6.641	7e-11	1.6e-06
Trametinib	GO:0098639 collagen binding involved ...	713	-1.500	-6.611	7.7e-11	1.7e-06
RDEA119	GO:0051045 negative regulation of mem...	668	-1.290	-6.616	7.8e-11	1.7e-06
NPK76-II-72-1	GO:0008380 RNA splicing	733	-1.572	-6.603	7.9e-11	1.7e-06
Lenalidomide	GO:0009451 RNA modification	677	-0.600	-6.601	8.5e-11	1.8e-06
PF-562271	GO:0005852 eukaryotic translation ini...	680	-0.772	-6.597	8.7e-11	1.8e-06
Bleomycin (50 uM)	GO:2000772 regulation of cellular sen...	746	-1.660	-6.583	8.9e-11	1.8e-06
GSK-650394	GO:0006412 translation	680	-1.873	-6.590	9.1e-11	1.9e-06
NPK76-II-72-1	GO:0006338 chromatin remodeling	733	-2.214	-6.572	9.7e-11	1.9e-06
Nutlin-3a	GO:0048515 spermatid differentiation	676	-0.914	-6.578	9.9e-11	2e-06
Bleomycin (50 uM)	GO:0005925 focal adhesion	746	-3.044	-6.558	1e-10	2.1e-06
PF-562271	GO:0003743 translation initiation fac...	680	-1.107	-6.556	1.1e-10	2.1e-06
RDEA119	GO:0097296 activation of cysteine-typ...	668	-1.424	-6.555	1.1e-10	2.1e-06
GSK-650394	GO:0006415 translational termination	680	-1.303	-6.558	1.1e-10	2.1e-06
GSK-650394	GO:0019058 viral life cycle	680	-1.867	-6.560	1.1e-10	2.1e-06
Lenalidomide	GO:0042645 mitochondrial nucleoid	677	-0.637	-6.553	1.2e-10	2.1e-06
NPK76-II-72-1	GO:0006397 mRNA processing	733	-1.791	-6.535	1.2e-10	2.2e-06
PF-562271	GO:0005840 ribosome	680	-1.099	-6.536	1.3e-10	2.3e-06
RDEA119	GO:2001238 positive regulation of ext...	668	-1.843	-6.518	1.5e-10	2.6e-06
RDEA119	GO:0005925 focal adhesion	668	-2.649	-6.516	1.5e-10	2.6e-06
Afatinib	GO:0030057 desmosome	675	-1.735	-6.511	1.5e-10	2.6e-06
BMN-673	GO:0001649 osteoblast differentiation	727	-3.093	-6.493	1.6e-10	2.8e-06
EX-527	GO:0060056 mammary gland involution	731	0.513	6.487	1.7e-10	2.8e-06
Trametinib	GO:0048008 platelet-derived growth fa...	713	-2.273	-6.473	1.8e-10	2.9e-06
GSK-650394	GO:0003735 structural constituent of ...	680	-1.588	-6.480	1.8e-10	2.9e-06
Trametinib	GO:2001238 positive regulation of ext...	713	-2.344	-6.477	1.8e-10	2.9e-06
YK 4-279	GO:0006412 translation	627	-1.493	-6.494	1.8e-10	2.9e-06
PF-562271	GO:0001731 formation of translation p...	680	-0.814	-6.485	1.8e-10	2.9e-06
Etoposide	GO:0030529 ribonucleoprotein complex	696	-2.454	-6.481	1.8e-10	2.9e-06
Bleomycin (50 uM)	GO:0033588 Elongator holoenzyme complex	746	-1.170	-6.468	1.8e-10	2.9e-06
KIN001-270	GO:0050792 regulation of viral process	733	0.497	6.468	1.9e-10	2.9e-06
Trametinib	GO:0048553 negative regulation of met...	713	-1.514	-6.463	2e-10	3.1e-06
NPK76-II-72-1	GO:0000398 mRNA splicing, via spliceo...	733	-1.444	-6.457	2e-10	3.1e-06
Vismodegib	GO:0006544 glycine metabolic process	676	-0.690	-6.463	2e-10	3.1e-06
Bleomycin (50 uM)	GO:0033209 tumor necrosis factor-medi...	746	-2.777	-6.452	2e-10	3.1e-06
RDEA119 (rescreen)	GO:0071438 invadopodium membrane	698	-1.028	-6.443	2.2e-10	3.3e-06
Afatinib (rescreen)	GO:0086073 bundle of His cell-Purkinj...	727	-1.160	-6.438	2.2e-10	3.3e-06
Afatinib (rescreen)	GO:0086083 cell adhesive protein bind...	727	-1.160	-6.438	2.2e-10	3.3e-06
Trametinib	GO:0097296 activation of cysteine-typ...	713	-1.748	-6.442	2.2e-10	3.3e-06
Trametinib	GO:0042127 regulation of cell prolife...	713	-4.053	-6.432	2.4e-10	3.4e-06
CCT018159	GO:0055086 nucleobase-containing smal...	715	-1.321	-6.417	2.6e-10	3.7e-06
Vismodegib	GO:0006144 purine nucleobase metaboli...	676	-0.945	-6.409	2.8e-10	4e-06
NPK76-II-72-1	GO:0003729 mRNA binding	733	-1.845	-6.395	2.9e-10	4.1e-06

Bleomycin (50 uM)	GO:0006446 regulation of translationa...	746	-1.669	-6.397	2.9e-10	4e-06
T0901317	GO:0001731 formation of translation p...	728	-0.607	-6.392	3e-10	4.1e-06
Nutlin-3a	GO:0060411 cardiac septum morphogenesis	676	-1.149	-6.389	3.2e-10	4.3e-06
RDEA119	GO:0070527 platelet aggregation	668	-1.963	-6.390	3.2e-10	4.3e-06
Bleomycin (50 uM)	GO:0010332 response to gamma radiation	746	-2.322	-6.377	3.2e-10	4.3e-06
Gemcitabine	GO:0019058 viral life cycle	685	-2.959	-6.386	3.2e-10	4.3e-06
Etoposide	GO:0006413 translational initiation	696	-1.775	-6.382	3.3e-10	4.3e-06
RDEA119	GO:0031581 hemidesmosome assembly	668	-1.346	-6.385	3.3e-10	4.3e-06
Lenalidomide	GO:0007005 mitochondrion organization	677	-1.030	-6.380	3.4e-10	4.3e-06
Nutlin-3a	GO:0032471 negative regulation of end...	676	-1.005	-6.381	3.4e-10	4.3e-06
RDEA119 (rescreen)	GO:0032587 ruffle membrane	698	-2.412	-6.377	3.4e-10	4.3e-06
Trametinib	GO:0070527 platelet aggregation	713	-2.525	-6.369	3.5e-10	4.4e-06
YK 4-279	GO:0003743 translation initiation fac...	627	-1.368	-6.381	3.5e-10	4.5e-06
Etoposide	GO:0005840 ribosome	696	-1.940	-6.367	3.6e-10	4.5e-06
PF-562271	GO:0002181 cytoplasmic translation	680	-0.726	-6.367	3.7e-10	4.6e-06
NPK76-II-72-1	GO:0015030 Cajal body	733	-1.339	-6.357	3.7e-10	4.6e-06
Gemcitabine	GO:0019083 viral transcription	685	-2.288	-6.363	3.7e-10	4.6e-06
Afatinib (rescreen)	GO:0030057 desmosome	727	-1.785	-6.346	4e-10	4.9e-06
RDEA119 (rescreen)	GO:0051045 negative regulation of mem...	698	-1.101	-6.340	4.2e-10	5.1e-06
NPK76-II-72-1	GO:0005840 ribosome	733	-1.363	-6.331	4.3e-10	5.1e-06
PF-562271	GO:0016282 eukaryotic 43S preinitiati...	680	-0.735	-6.339	4.3e-10	5.1e-06
Nutlin-3a	GO:0045869 negative regulation of sin...	676	-0.827	-6.339	4.4e-10	5.1e-06
Epothilone B	GO:0019083 viral transcription	689	-1.583	-6.324	4.7e-10	5.5e-06
YK 4-279	GO:0019083 viral transcription	627	-1.136	-6.333	4.7e-10	5.5e-06
PF-562271	GO:0006614 SRP-dependent cotranslatio...	680	-0.956	-6.318	4.9e-10	5.7e-06
Bleomycin (50 uM)	GO:0090090 negative regulation of can...	746	-3.257	-6.305	5e-10	5.8e-06
PF-562271	GO:0071541 eukaryotic translation ini...	680	-0.659	-6.312	5.1e-10	5.8e-06
GSK-650394	GO:0006414 translational elongation	680	-1.321	-6.306	5.3e-10	6e-06
JNK-9L	GO:0002181 cytoplasmic translation	695	-0.729	-6.299	5.4e-10	6.1e-06
PF-562271	GO:0019083 viral transcription	680	-0.907	-6.304	5.4e-10	6e-06
Trametinib	GO:0008305 integrin complex	713	-2.408	-6.294	5.5e-10	6.1e-06
RDEA119 (rescreen)	GO:0048553 negative regulation of met...	698	-1.019	-6.295	5.6e-10	6.1e-06
NPK76-II-72-1	GO:0000027 ribosomal large subunit as...	733	-1.002	-6.290	5.6e-10	6.1e-06
Nutlin-3a	GO:0000780 condensed nuclear chromoso...	676	0.878	6.293	5.7e-10	6.2e-06
YK 4-279	GO:0009168 purine ribonucleoside mono...	627	-1.058	-6.300	5.8e-10	6.2e-06
YK 4-279	GO:0071541 eukaryotic translation ini...	627	-0.836	-6.297	5.9e-10	6.3e-06

Associations for Reactome

Table A3: Reactome vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
WZ3105	Abacavir transmembrane transport	734	1.268	8.505	1.1e-16	2.2e-11
RDEA119 (rescreen)	Dissolution of Fibrin Clot	698	-1.570	-8.501	1.2e-16	2.2e-11
Trametinib	Dissolution of Fibrin Clot	713	-2.277	-8.428	2.1e-16	2.5e-11
Trametinib	Laminin interactions	713	-2.534	-8.138	1.9e-15	1.5e-10
RDEA119	Dissolution of Fibrin Clot	668	-1.710	-8.143	2e-15	1.5e-10
CCT018159	Purine metabolism	715	-1.343	-7.503	1.9e-13	1.1e-08
RDEA119 (rescreen)	Laminin interactions	698	-1.639	-7.494	2.1e-13	1.1e-08
Trametinib	Type I hemidesmosome assembly	713	-1.785	-7.302	7.9e-13	3.6e-08
Etoposide	Influenza Life Cycle	696	-2.076	-7.234	1.3e-12	5.2e-08
Etoposide	Influenza Infection	696	-2.120	-7.219	1.4e-12	5.2e-08
Etoposide	Influenza Viral RNA Transcription and...	696	-1.999	-7.111	3e-12	9.9e-08
GSK-650394	Cap-dependent Translation Initiation	680	-1.477	-6.997	6.5e-12	1.8e-07
GSK-650394	Eukaryotic Translation Initiation	680	-1.477	-6.997	6.5e-12	1.8e-07
GSK-650394	Influenza Viral RNA Transcription and...	680	-1.672	-6.971	7.8e-12	2e-07
RDEA119	Laminin interactions	668	-1.728	-6.921	1.1e-11	2.6e-07
GSK-650394	Activation of the mRNA upon binding o...	680	-1.493	-6.911	1.2e-11	2.6e-07
Trametinib	Degradation of the extracellular matrix	713	-3.332	-6.896	1.2e-11	2.6e-07
Gemcitabine	Influenza Infection	685	-2.808	-6.892	1.3e-11	2.6e-07
EHT 1864	Mitochondrial tRNA aminoacylation	737	-0.737	-6.850	1.6e-11	3.1e-07
GSK-650394	Influenza Life Cycle	680	-1.683	-6.854	1.7e-11	3.1e-07
Gemcitabine	Influenza Life Cycle	685	-2.713	-6.805	2.3e-11	4e-07
RDEA119 (rescreen)	Degradation of the extracellular matrix	698	-2.253	-6.790	2.5e-11	4.1e-07
PF-562271	Influenza Life Cycle	680	-1.079	-6.783	2.6e-11	4.1e-07
PF-562271	Translation	680	-1.090	-6.781	2.7e-11	4.1e-07
GSK-650394	Influenza Infection	680	-1.703	-6.775	2.8e-11	4.1e-07
GSK-650394	GTP hydrolysis and joining of the 60S...	680	-1.374	-6.753	3.2e-11	4.1e-07
Bleomycin (50 uM)	Insulin-like Growth Factor-2 mRNA Bin...	746	-1.399	-6.745	3.2e-11	4.1e-07
GSK-650394	Ribosomal scanning and start codon re...	680	-1.461	-6.748	3.3e-11	4.1e-07
GSK-650394	Translation	680	-1.675	-6.751	3.3e-11	4.1e-07
RDEA119 (rescreen)	Type I hemidesmosome assembly	698	-1.144	-6.743	3.4e-11	4.1e-07

GSK-650394	Formation of the ternary complex, and...	680	-1.408	-6.730	3.7e-11	4.3e-07
GSK-650394	Translation initiation complex formation	680	-1.444	-6.729	3.7e-11	4.3e-07
GSK-650394	L13a-mediated translational silencing...	680	-1.345	-6.695	4.7e-11	5e-07
GSK-650394	3' -UTR-mediated translational regula...	680	-1.345	-6.695	4.7e-11	5e-07
PF-562271	Ribosomal scanning and start codon re...	680	-0.936	-6.675	5.3e-11	5.3e-07
GSK-650394	Nonsense Mediated Decay (NMD) enhance...	680	-1.483	-6.672	5.4e-11	5.3e-07
GSK-650394	Nonsense-Mediated Decay (NMD)	680	-1.483	-6.672	5.4e-11	5.3e-07
PF-562271	Cap-dependent Translation Initiation	680	-0.912	-6.650	6.2e-11	5.8e-07
PF-562271	Eukaryotic Translation Initiation	680	-0.912	-6.650	6.2e-11	5.8e-07
PF-562271	Influenza Infection	680	-1.081	-6.641	6.6e-11	6e-07
NPK76-II-72-1	Influenza Life Cycle	733	-1.368	-6.615	7.4e-11	6.5e-07
Trametinib	Insulin-like Growth Factor-2 mRNA Bin...	713	-1.535	-6.587	8.9e-11	7.3e-07
CCT018159	Purine ribonucleoside monophosphate b...	715	-0.714	-6.587	8.9e-11	7.3e-07
PF-562271	Activation of the mRNA upon binding o...	680	-0.925	-6.592	9e-11	7.3e-07
Gemcitabine	Influenza Viral RNA Transcription and...	685	-2.574	-6.588	9.2e-11	7.3e-07
Trametinib	Uptake and function of anthrax toxins	713	-1.833	-6.582	9.3e-11	7.3e-07
PF-562271	GTP hydrolysis and joining of the 60S...	680	-0.864	-6.544	1.2e-10	9.2e-07
NPK76-II-72-1	Influenza Infection	733	-1.387	-6.540	1.2e-10	9.2e-07
GSK-650394	Formation of a pool of free 40S subunits	680	-1.278	-6.533	1.3e-10	9.7e-07
PF-562271	Influenza Viral RNA Transcription and...	680	-1.015	-6.508	1.5e-10	1.1e-06
PF-562271	Translation initiation complex formation	680	-0.906	-6.512	1.5e-10	1.1e-06
RDEA119	Degradation of the extracellular matrix	668	-2.479	-6.516	1.5e-10	1.1e-06
Thapsigargin	Influenza Life Cycle	680	-1.961	-6.446	2.2e-10	1.5e-06
PF-562271	L13a-mediated translational silencing...	680	-0.836	-6.414	2.7e-10	1.8e-06
GSK-650394	Nonsense Mediated Decay (NMD) indepen...	680	-1.265	-6.415	2.7e-10	1.8e-06
PF-562271	3' -UTR-mediated translational regula...	680	-0.836	-6.414	2.7e-10	1.8e-06
PF-562271	Formation of the ternary complex, and...	680	-0.870	-6.408	2.8e-10	1.8e-06
Thapsigargin	Influenza Infection	680	-1.996	-6.413	2.8e-10	1.8e-06
Trametinib	Non-integrin membrane-ECM interactions	713	-2.349	-6.391	3.1e-10	1.9e-06
GSK-650394	Eukaryotic Translation Termination	680	-1.229	-6.391	3.2e-10	1.9e-06
Epothilone B	Influenza Life Cycle	689	-1.777	-6.376	3.4e-10	2e-06
GSK-650394	Eukaryotic Translation Elongation	680	-1.204	-6.372	3.5e-10	2.1e-06
Epothilone B	Influenza Infection	689	-1.814	-6.362	3.7e-10	2.2e-06
GSK-650394	Peptide chain elongation	680	-1.185	-6.353	4e-10	2.3e-06
WZ3105	Abacavir transport and metabolism	734	1.299	6.342	4.1e-10	2.3e-06
GSK-650394	Viral mRNA Translation	680	-1.223	-6.332	4.5e-10	2.5e-06
RDEA119 (rescreen)	TRAIL signaling	698	-0.973	-6.328	4.6e-10	2.5e-06
Trametinib	TRAIL signaling	713	-1.418	-6.320	4.7e-10	2.5e-06
YK 4-279	Influenza Life Cycle	627	-1.274	-6.320	5.1e-10	2.7e-06
NPK76-II-72-1	Influenza Viral RNA Transcription and...	733	-1.278	-6.301	5.2e-10	2.7e-06
Trametinib	Purine catabolism	713	-1.775	-6.302	5.3e-10	2.7e-06
Lenalidomide	Metabolism of non-coding RNA	677	-0.547	-6.300	5.5e-10	2.7e-06
Lenalidomide	snRNP Assembly	677	-0.547	-6.300	5.5e-10	2.7e-06
Epothilone B	Influenza Viral RNA Transcription and...	689	-1.711	-6.274	6.4e-10	3.1e-06
PF-562271	Formation of a pool of free 40S subunits	680	-0.796	-6.268	6.7e-10	3.2e-06
Bleomycin (50 uM)	Neurotransmitter Release Cycle	746	2.439	6.252	7e-10	3.3e-06
T0901317	Cap-dependent Translation Initiation	728	-0.655	-6.251	7.1e-10	3.3e-06
T0901317	Eukaryotic Translation Initiation	728	-0.655	-6.251	7.1e-10	3.3e-06
(5Z)-7-Oxozeaenol	Loss of Function of TGFBR1 in Cancer	729	-0.921	-6.238	7.7e-10	3.5e-06
(5Z)-7-Oxozeaenol	TGFBR1 KD Mutants in Cancer	729	-0.921	-6.238	7.7e-10	3.5e-06
Thapsigargin	Influenza Viral RNA Transcription and...	680	-1.860	-6.240	7.9e-10	3.5e-06
YK 4-279	Influenza Infection	627	-1.288	-6.240	8.3e-10	3.7e-06
Bleomycin (50 uM)	Dissolution of Fibrin Clot	746	-1.560	-6.220	8.4e-10	3.7e-06
Afatinib	Apoptotic cleavage of cell adhesion ...	675	-1.278	-6.228	8.5e-10	3.7e-06
YK 4-279	Translation	627	-1.252	-6.233	8.7e-10	3.7e-06
PF-562271	SRP-dependent cotranslational protein...	680	-0.951	-6.217	9.1e-10	3.8e-06
RDEA119 (rescreen)	Non-integrin membrane-ECM interactions	698	-1.575	-6.204	9.7e-10	4e-06
YK 4-279	Influenza Viral RNA Transcription and...	627	-1.222	-6.198	1.1e-09	4.3e-06
RDEA119 (rescreen)	Insulin-like Growth Factor-2 mRNA Bin...	698	-0.994	-6.190	1.1e-09	4.3e-06
Trametinib	Syndecan interactions	713	-1.886	-6.184	1.1e-09	4.3e-06
Lenalidomide	Interactions of Rev with host cellula...	677	-0.520	-6.174	1.2e-09	4.7e-06
Lenalidomide	Nuclear import of Rev protein	677	-0.530	-6.168	1.2e-09	4.8e-06
Lenalidomide	Rev-mediated nuclear export of HIV RNA	677	-0.519	-6.155	1.3e-09	5.2e-06
Etoposide	Folding of actin by CCT/TriC	696	-1.031	-6.140	1.4e-09	5.4e-06
Etoposide	Nonsense Mediated Decay (NMD) enhance...	696	-1.607	-6.139	1.4e-09	5.4e-06
Etoposide	Nonsense-Mediated Decay (NMD)	696	-1.607	-6.139	1.4e-09	5.4e-06
RDEA119	Type I hemidesmosome assembly	668	-1.211	-6.147	1.4e-09	5.4e-06
Vismodegib	Purine metabolism	676	-0.876	-6.129	1.5e-09	5.6e-06
Epothilone B	Ribosomal scanning and start codon re...	689	-1.506	-6.127	1.5e-09	5.6e-06
Etoposide	Cap-dependent Translation Initiation	696	-1.528	-6.128	1.5e-09	5.6e-06
Etoposide	Eukaryotic Translation Initiation	696	-1.528	-6.128	1.5e-09	5.6e-06
Bleomycin (50 uM)	Infectious disease	746	-2.675	-6.104	1.7e-09	6e-06
PF-562271	Nonsense Mediated Decay (NMD) indepen...	680	-0.784	-6.116	1.7e-09	5.9e-06
GSK-650394	SRP-dependent cotranslational protein...	680	-1.439	-6.110	1.7e-09	6e-06
Obatoclax Mesylate	Influenza Infection	684	-1.608	-6.103	1.8e-09	6.1e-06

Lenalidomide	Viral Messenger RNA Synthesis	677	-0.589	-6.104	1.8e-09	6.1e-06
Lenalidomide	Glucose transport	677	-0.704	-6.098	1.9e-09	6.3e-06
Nutlin-3a	Mitochondrial iron-sulfur cluster bio...	676	-0.911	-6.090	1.9e-09	6.4e-06
Epothilone B	Cap-dependent Translation Initiation	689	-1.463	-6.093	1.9e-09	6.3e-06
Epothilone B	Eukaryotic Translation Initiation	689	-1.463	-6.093	1.9e-09	6.3e-06
Epothilone B	Activation of the mRNA upon binding o...	689	-1.496	-6.084	2e-09	6.5e-06
CCT018159	Metabolism of nucleotides	715	-1.198	-6.081	2e-09	6.5e-06
Etoposide	Ribosomal scanning and start codon re...	696	-1.548	-6.060	2.3e-09	7.3e-06
YM155	Biological oxidations	706	4.013	6.058	2.3e-09	7.3e-06
YK 4-279	Cap-dependent Translation Initiation	627	-1.034	-6.052	2.5e-09	7.9e-06
YK 4-279	Eukaryotic Translation Initiation	627	-1.034	-6.052	2.5e-09	7.9e-06
Bleomycin (50 uM)	Interleukin-6 signaling	746	-1.361	-6.032	2.6e-09	8e-06
Lenalidomide	Transport of Mature mRNAs Derived fro...	677	-0.555	-6.042	2.6e-09	8e-06
RDEA119 (rescreen)	Ligand-dependent caspase activation	698	-1.113	-6.033	2.7e-09	8.2e-06
Obatoclax Mesylate	Influenza Life Cycle	684	-1.553	-6.024	2.8e-09	8.4e-06
PF-562271	Viral mRNA Translation	680	-0.757	-6.029	2.8e-09	8.4e-06
Epothilone B	Formation of the ternary complex, and...	689	-1.430	-6.022	2.9e-09	8.4e-06
XMD13-2	Nonsense Mediated Decay (NMD) enhance...	733	-0.891	-6.015	2.9e-09	8.4e-06
XMD13-2	Nonsense-Mediated Decay (NMD)	733	-0.891	-6.015	2.9e-09	8.4e-06
Lenalidomide	Transport of Ribonucleoproteins into ...	677	-0.528	-6.022	2.9e-09	8.4e-06
T0901317	L13a-mediated translational silencing...	728	-0.600	-6.008	3e-09	8.4e-06
RDEA119	Ligand-dependent caspase activation	668	-1.293	-6.017	3e-09	8.4e-06
PF-562271	Nonsense Mediated Decay (NMD) enhance...	680	-0.874	-6.017	3e-09	8.4e-06
PF-562271	Nonsense-Mediated Decay (NMD)	680	-0.874	-6.017	3e-09	8.4e-06
PF-562271	Peptide chain elongation	680	-0.730	-6.017	3e-09	8.4e-06
T0901317	3' -UTR-mediated translational regula...	728	-0.600	-6.008	3e-09	8.4e-06
Lenalidomide	Metabolism of nucleotides	677	-0.843	-6.008	3.1e-09	8.6e-06
Epothilone B	Translation initiation complex formation	689	-1.465	-6.008	3.1e-09	8.6e-06
Etoposide	Translation	696	-1.757	-6.001	3.2e-09	8.6e-06
Lenalidomide	Transport of Mature mRNA Derived from...	677	-0.553	-6.005	3.2e-09	8.6e-06
T0901317	GTP hydrolysis and joining of the 60S...	728	-0.606	-5.992	3.3e-09	8.6e-06
Gemcitabine	Nonsense Mediated Decay (NMD) enhance...	685	-2.172	-5.997	3.3e-09	8.6e-06
Tubastatin A	Nonsense Mediated Decay (NMD) enhance...	731	-0.862	-5.993	3.3e-09	8.6e-06
Gemcitabine	Nonsense-Mediated Decay (NMD)	685	-2.172	-5.997	3.3e-09	8.6e-06
Tubastatin A	Nonsense-Mediated Decay (NMD)	731	-0.862	-5.993	3.3e-09	8.6e-06
YK 4-279	Synthesis of diphthamide-EEF2	627	-0.756	-6.001	3.4e-09	8.7e-06
Epothilone B	tRNA Aminoacylation	689	-1.482	-5.993	3.4e-09	8.7e-06
PF-562271	Eukaryotic Translation Elongation	680	-0.737	-5.995	3.4e-09	8.7e-06
Lenalidomide	Export of Viral Ribonucleoproteins fr...	677	-0.525	-5.993	3.4e-09	8.7e-06
Lenalidomide	Transport of the SLBP independent Mat...	677	-0.537	-5.990	3.5e-09	8.8e-06
Etoposide	GTP hydrolysis and joining of the 60S...	696	-1.437	-5.984	3.6e-09	8.9e-06
Lenalidomide	Transport of the SLBP Dependant Matur...	677	-0.536	-5.985	3.6e-09	8.9e-06
Obatoclax Mesylate	Influenza Viral RNA Transcription and...	684	-1.507	-5.975	3.8e-09	9.1e-06
Gemcitabine	Cap-dependent Translation Initiation	685	-2.065	-5.976	3.8e-09	9.1e-06
Gemcitabine	Eukaryotic Translation Initiation	685	-2.065	-5.976	3.8e-09	9.1e-06

A.2 *SPEED platform*

Table A4: Optimised SPEED scores vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
RDEA119	MAPK_PI3K	668	-0.326	-6.321	4.9e-10	1.1e-06
Trametinib	MAPK_PI3K	713	-0.409	-6.244	7.5e-10	1.1e-06
RDEA119 (rescreen)	MAPK_PI3K	698	-0.262	-5.856	7.4e-09	7.2e-06
Bleomycin (50 uM)	VEGF	746	-0.256	-5.249	2e-07	0.00012
PD-0325901	MAPK_PI3K	669	-0.276	-5.249	2.1e-07	0.00012
Afatinib (rescreen)	PI3K_only	727	-0.206	-4.737	2.6e-06	0.0013
MLN4924	Wnt	553	-0.173	-4.719	3e-06	0.0013
XAV 939	TNFa	730	-0.117	-4.632	4.3e-06	0.0016
Axitinib	Wnt	673	-0.116	-4.491	8.4e-06	0.0027
Afatinib	PI3K_only	675	-0.179	-4.443	1e-05	0.003
Gefitinib	PI3K_only	673	-0.130	-4.379	1.4e-05	0.0037
QS11	Wnt	689	-0.135	-4.354	1.5e-05	0.0038
Etoposide	Wnt	696	-0.186	-4.327	1.7e-05	0.0039
Elesclomol	VEGF	675	-0.204	-4.279	2.2e-05	0.0045
Bleomycin	TNFa	682	-0.304	-4.257	2.4e-05	0.0046
GSK-650394	Wnt	680	-0.156	-4.226	2.7e-05	0.005
AG-014699	TGFB	736	-0.096	-4.209	2.9e-05	0.005
Bleomycin (50 uM)	TNFa	746	-0.223	-4.156	3.6e-05	0.0059
CP724714	H2O2	734	0.128	4.099	4.6e-05	0.0065
Cisplatin	Wnt	677	-0.108	-4.103	4.6e-05	0.0065
Thapsigargin	Wnt	680	-0.186	-4.097	4.7e-05	0.0065

BX-912	TGFB	733	-0.143	-4.078	5.1e-05	0.0067
EX-527	Wnt	731	-0.055	-4.041	5.9e-05	0.0075
BMN-673	VEGF	727	-0.185	-4.005	6.9e-05	0.0081
HG-6-64-1	H2O2	691	-0.212	-4.003	7e-05	0.0081
TW 37	VEGF	735	-0.123	-3.984	7.5e-05	0.0082
ABT-888	Wnt	676	-0.064	-3.984	7.6e-05	0.0082
Midostaurin	IL-1	696	-0.131	-3.958	8.4e-05	0.0083
Afatinib	H2O2	675	0.189	3.953	8.6e-05	0.0083
Axitinib	TGFB	673	-0.114	-3.954	8.6e-05	0.0083
EHT 1864	MAPK_PI3K	737	0.112	3.945	8.8e-05	0.0083
Gemcitabine	Wnt	685	-0.232	-3.884	0.00011	0.0099
Lenalidomide	Wnt	677	-0.065	-3.899	0.00011	0.0097
Pazopanib	TGFB	688	-0.121	-3.864	0.00012	0.01
GW 441756	Wnt	674	-0.103	-3.878	0.00012	0.0099
Camptothecin	TLR	675	-0.200	-3.842	0.00013	0.011
SB52334	H2O2	732	0.150	3.802	0.00016	0.012
CI-1040	MAPK_PI3K	669	-0.168	-3.779	0.00017	0.013
RDEA119 (rescreen)	TNFa	698	-0.153	-3.783	0.00017	0.013
Bleomycin	TLR	682	-0.280	-3.769	0.00018	0.013
Midostaurin	H2O2	696	-0.156	-3.737	0.0002	0.014
piperlongumine	H2O2	741	-0.113	-3.731	0.00021	0.014
YK 4-279	Wnt	627	-0.112	-3.735	0.00021	0.014
Docetaxel	H2O2	676	-0.170	-3.699	0.00023	0.016
Trametinib	TNFa	713	-0.218	-3.686	0.00025	0.016
FH535	PI3K_only	686	-0.130	-3.661	0.00027	0.017
RO-3306	Wnt	676	-0.089	-3.661	0.00027	0.017
Vinorelbine	Wnt	696	-0.134	-3.657	0.00028	0.017
XL-880	VEGF	730	-0.128	-3.641	0.00029	0.017
RDEA119	TNFa	668	-0.172	-3.637	0.0003	0.017
GSK690693	MAPK_PI3K	732	0.180	3.607	0.00033	0.019
XAV 939	TLR	730	-0.095	-3.594	0.00035	0.02
Afatinib	TGFB	675	0.131	3.589	0.00036	0.02
rTRAIL	Wnt	731	-0.110	-3.572	0.00038	0.02
Bleomycin (50 uM)	TLR	746	-0.199	-3.559	0.0004	0.02
AZD6482	TNFa	691	-0.139	-3.560	0.0004	0.02
AZD6482	TNFa	691	-0.139	-3.560	0.0004	0.02
NVP-BEZ235	H2O2	670	-0.144	-3.545	0.00042	0.021
Bleomycin (50 uM)	IL-1	746	-0.176	-3.534	0.00043	0.021
(5Z)-7-Oxozeaenol	MAPK_PI3K	729	-0.147	-3.540	0.00043	0.021
RDEA119	TLR	668	-0.171	-3.525	0.00045	0.022
Afatinib (rescreen)	TGFB	727	0.137	3.504	0.00049	0.023
Docetaxel	TNFa	676	-0.150	-3.501	0.00049	0.023
Bleomycin (50 uM)	H2O2	746	-0.211	-3.471	0.00055	0.025
HG-6-64-1	IL-1	691	-0.147	-3.470	0.00055	0.025
Camptothecin	TNFa	675	-0.177	-3.465	0.00056	0.025
Vinblastine	Wnt	677	-0.113	-3.449	0.0006	0.026
piperlongumine	VEGF	741	-0.085	-3.430	0.00064	0.027
Trametinib	IL-1	713	-0.188	-3.421	0.00066	0.027
AZD6244	MAPK_PI3K	658	-0.179	-3.423	0.00066	0.027
AZD6244	MAPK_PI3K	658	-0.179	-3.423	0.00066	0.027
BIRB 0796	Wnt	670	-0.073	-3.420	0.00066	0.027
PLX4720 (rescreen)	TNFa	730	-0.105	-3.405	0.0007	0.028
AS601245	TNFa	686	-0.119	-3.387	0.00075	0.029
RDEA119 (rescreen)	IL-1	698	-0.127	-3.359	0.00083	0.032
Elesclomol	Wnt	675	-0.133	-3.354	0.00084	0.032
TGX221	TGFB	290	-0.152	-3.371	0.00086	0.032
BX-795	VEGF	676	-0.111	-3.346	0.00087	0.032
MP470	H2O2	730	0.171	3.328	0.00092	0.034
RDEA119 (rescreen)	TLR	698	-0.142	-3.311	0.00098	0.035
SB-505124	Wnt	741	-0.062	-3.309	0.00098	0.035
Epothilone B	Wnt	689	-0.137	-3.295	0.001	0.037
BMN-673	Wnt	727	-0.123	-3.247	0.0012	0.043
Docetaxel	IL-1	676	-0.125	-3.240	0.0013	0.044
(5Z)-7-Oxozeaenol	VEGF	729	-0.110	-3.215	0.0014	0.047
BMS-345541	H2O2	734	0.110	3.186	0.0015	0.049
Cisplatin	VEGF	677	-0.102	-3.197	0.0015	0.049
Docetaxel	Wnt	676	-0.100	-3.191	0.0015	0.049
NU-7441	Wnt	674	-0.068	-3.196	0.0015	0.049
JNK Inhibitor VIII	Wnt	674	-0.059	-3.180	0.0015	0.049
DMOG	H2O2	692	-0.145	-3.177	0.0016	0.049
VNLG/124	JAK-STAT	731	-0.079	-3.175	0.0016	0.049

B ASSOCIATIONS FOR EVALUATING SIGNATURES (CHAPTER 4)

B.1 *Pathway scores and mutations*

Table B1: Gene Ontology vs. mutations (pan-cancer)

Mutation	Pathway	Size	Effect	Wald stat.	P-value	FDR
TP53	NFkB	1339	0.086	7.784	9.1e-15	4.6e-11
BRAF	Trail	313	0.202	7.379	2e-13	5e-10
BRAF	MAPK	313	0.051	7.199	7.3e-13	1.2e-09
BRAF	JAK-STAT	313	0.112	6.960	4e-12	5.1e-09
NRAS	JAK-STAT	56	-0.235	-6.371	2.1e-10	1.8e-07
CDH1	NFkB	170	-0.163	-6.371	2.1e-10	1.8e-07
TP53	PI3K	1339	-0.049	-5.870	4.7e-09	3.5e-06
TP53	p53	1339	-0.050	-5.797	7.3e-09	4.7e-06
TP53	TNFa	1339	0.050	5.677	1.5e-08	8.4e-06
BRAF	PI3K	313	0.081	5.640	1.8e-08	9.3e-06
BRAF	EGFR	313	0.129	5.098	3.6e-07	0.00017
NRAS	Trail	56	-0.311	-4.961	7.3e-07	0.00031
CASP8	TNFa	88	0.133	4.793	1.7e-06	0.00067
ZFHX3	TNFa	166	0.096	4.720	2.4e-06	0.00089
TP53	MAPK	1339	0.019	4.703	2.7e-06	0.0009
BRAF	p53	313	0.069	4.678	3e-06	0.00096
CASP8	JAK-STAT	88	0.138	4.658	3.3e-06	0.00099
BCL11A	NFkB	68	0.185	4.635	3.7e-06	0.001
BCL11A	TNFa	68	0.143	4.558	5.3e-06	0.0014
ZFHX3	NFkB	166	0.118	4.531	6e-06	0.0015
NRAS	PI3K	56	-0.148	-4.524	6.3e-06	0.0015
ARID1A	p53	318	0.067	4.516	6.5e-06	0.0015
MLH3	NFkB	59	0.193	4.489	7.4e-06	0.0016
BCOR	TNFa	97	0.117	4.438	9.4e-06	0.0019
AHNAK	TNFa	195	0.084	4.433	9.5e-06	0.0019
EP300	p53	123	0.102	4.405	1.1e-05	0.0021
ARHGAP26	TNFa	41	0.178	4.409	1.1e-05	0.0021
ARFGEF2	p53	85	0.121	4.375	1.2e-05	0.0022
MLL2	NFkB	278	0.089	4.367	1.3e-05	0.0022
PIK3CA	PI3K	667	0.045	4.361	1.3e-05	0.0022
SETD2	TNFa	161	0.090	4.363	1.3e-05	0.0022
NRAS	MAPK	56	-0.070	-4.320	1.6e-05	0.0026
CASP8	MAPK	88	0.056	4.311	1.7e-05	0.0026
MLH3	TNFa	59	0.145	4.296	1.8e-05	0.0027
EP300	TNFa	123	0.101	4.279	1.9e-05	0.0028
GATA3	p53	136	0.094	4.258	2.1e-05	0.0029
ARFGEF1	TNFa	90	0.117	4.256	2.1e-05	0.0029
CDH1	p53	170	-0.084	-4.249	2.2e-05	0.003
AHNAK	NFkB	195	0.102	4.236	2.3e-05	0.0031
CDH1	PI3K	170	0.081	4.225	2.4e-05	0.0031
ARFGEF2	TNFa	85	0.118	4.181	3e-05	0.0037
MLL2	TNFa	278	0.067	4.170	3.1e-05	0.0038
ARHGAP26	NFkB	41	0.211	4.114	4e-05	0.0047
NOTCH1	MAPK	141	0.042	4.067	4.9e-05	0.0056
PTEN	EGFR	284	-0.108	-4.064	4.9e-05	0.0056
PBRM1	p53	206	0.073	4.058	5e-05	0.0056
KRAS	TGFb	309	0.041	4.047	5.3e-05	0.0056
SHMT1	TNFa	19	0.239	4.046	5.3e-05	0.0056
GATA3	Hypoxia	136	-0.049	-4.042	5.4e-05	0.0056
MMP2	NFkB	42	0.204	4.019	5.9e-05	0.0061
B2M	TNFa	34	0.178	4.014	6.1e-05	0.0061
CHD3	TNFa	107	0.101	4.009	6.2e-05	0.0061
G3BP1	TNFa	36	0.171	3.981	7e-05	0.0067
MLL3	NFkB	296	0.079	3.976	7.1e-05	0.0067
ARFGEF1	p53	90	0.106	3.966	7.4e-05	0.0069
BAP1	p53	89	0.107	3.958	7.7e-05	0.007
CTNNB1	JAK-STAT	182	-0.082	-3.943	8.2e-05	0.0073
PLCG1	TNFa	61	0.129	3.900	9.8e-05	0.0086
IRS2	Hypoxia	24	-0.111	-3.882	0.00011	0.009
KALRN	p53	152	0.081	3.869	0.00011	0.0091
MMP2	TNFa	42	0.155	3.871	0.00011	0.0091
ASH1L	NFkB	127	0.115	3.882	0.00011	0.009
PBRM1	TNFa	206	0.071	3.859	0.00012	0.0094
BCOR	NFkB	97	0.129	3.824	0.00013	0.01

CDH1	TNFa	170	-0.078	-3.838	0.00013	0.01
PTEN	p53	284	0.059	3.781	0.00016	0.012
EPHA2	NFkB	60	0.160	3.760	0.00017	0.013
MTOR	TNFa	136	0.084	3.758	0.00017	0.013
BRWD1	TNFa	105	0.096	3.751	0.00018	0.013
MLL2	p53	278	0.059	3.733	0.00019	0.014
NRAS	p53	56	-0.126	-3.732	0.00019	0.014
SF3A3	Hypoxia	22	0.112	3.732	0.00019	0.014
SETD2	NFkB	161	0.098	3.726	0.0002	0.014
SHMT1	NFkB	19	0.280	3.726	0.0002	0.014
VHL	VEGF	209	0.116	3.705	0.00021	0.014
ACSL6	TNFa	37	0.158	3.714	0.00021	0.014
ASH1L	TNFa	127	0.086	3.698	0.00022	0.014
ZNF292	TNFa	112	0.091	3.703	0.00022	0.014
ELF1	TNFa	36	0.158	3.676	0.00024	0.016
PLCG1	NFkB	61	0.154	3.655	0.00026	0.017
PTEN	TNFa	284	0.058	3.651	0.00026	0.017
ASH1L	p53	127	0.083	3.642	0.00027	0.017
ZNF292	NFkB	112	0.114	3.646	0.00027	0.017
PIK3CA	Hypoxia	667	-0.022	-3.636	0.00028	0.017
TCF12	p53	39	0.147	3.638	0.00028	0.017
CHD6	TNFa	127	0.084	3.611	0.00031	0.018
ARFGEF1	NFkB	90	0.126	3.612	0.00031	0.018
CNOT4	NFkB	40	0.187	3.598	0.00033	0.019
FGFR1	TGFb	34	-0.106	-3.590	0.00033	0.019
MYH14	TNFa	76	0.107	3.585	0.00034	0.019
MTOR	NFkB	136	0.102	3.576	0.00035	0.019
NRAS	EGFR	56	-0.207	-3.579	0.00035	0.019
PGR	NFkB	59	0.154	3.582	0.00035	0.019
VHL	p53	209	0.064	3.580	0.00035	0.019
ARID1B	NFkB	115	0.110	3.562	0.00037	0.02
KEAP1	VEGF	114	-0.149	-3.558	0.00038	0.02
EP300	NFkB	123	0.106	3.548	0.00039	0.021
CTNNB1	PI3K	182	-0.066	-3.543	0.0004	0.021
ELF1	NFkB	36	0.194	3.541	0.0004	0.021
RASA1	TNFa	87	0.099	3.546	0.0004	0.021
SOS2	TNFa	66	0.113	3.528	0.00042	0.021
KRAS	TNFa	309	0.054	3.524	0.00043	0.021
VHL	Hypoxia	209	0.035	3.527	0.00043	0.021
WNK1	MAPK	94	0.044	3.506	0.00046	0.023
CTNNB1	TGFb	182	-0.046	-3.498	0.00048	0.023
KRAS	PI3K	309	0.050	3.481	0.0005	0.024
G3BP1	NFkB	36	0.191	3.481	0.00051	0.024
NFE2L2	NFkB	105	0.112	3.461	0.00054	0.026
FBXO11	TNFa	41	0.139	3.448	0.00057	0.026
SVEP1	MAPK	156	0.034	3.447	0.00057	0.026
NF1	NFkB	162	0.091	3.446	0.00058	0.026
TP53BP1	TNFa	87	0.096	3.437	0.0006	0.027
ARFGEF2	NFkB	85	0.123	3.428	0.00061	0.028
GATA3	JAK-STAT	136	-0.082	-3.427	0.00062	0.028
MECOM	TGFb	76	-0.068	-3.404	0.00067	0.03
BPTF	TNFa	106	0.086	3.403	0.00067	0.03
PIK3CA	p53	667	0.037	3.396	0.00069	0.03
FGFR1	TNFa	34	0.150	3.393	0.0007	0.03
CASP8	EGFR	88	0.157	3.390	0.00071	0.03
HLA-A	TNFa	54	0.119	3.380	0.00073	0.03
SVEP1	TNFa	156	0.071	3.382	0.00073	0.03
XRN1	TGFb	79	-0.066	-3.380	0.00073	0.03
CNOT4	TNFa	40	0.138	3.378	0.00074	0.03
PBRM1	EGFR	206	0.104	3.377	0.00074	0.03
SOS2	NFkB	66	0.137	3.376	0.00074	0.03
SPTAN1	TNFa	111	0.084	3.371	0.00076	0.031
GATA3	MAPK	136	-0.036	-3.361	0.00079	0.032
NR4A2	NFkB	47	0.161	3.357	0.0008	0.032
BRAF	TNFa	313	0.051	3.353	0.00081	0.032
NCK1	TNFa	23	0.180	3.337	0.00085	0.033
TAF1	TNFa	111	0.083	3.337	0.00086	0.033
ABL2	TNFa	54	0.117	3.322	0.0009	0.035
ERCC2	NFkB	41	0.171	3.319	0.00091	0.035
RBM10	TNFa	59	0.112	3.321	0.00091	0.035
FAF1	TNFa	37	0.141	3.314	0.00093	0.035
BAP1	TGFb	89	-0.061	-3.312	0.00093	0.035
DHX15	NFkB	36	0.182	3.311	0.00094	0.035
FBXW7	NFkB	134	0.095	3.310	0.00094	0.035
KDM5C	TNFa	78	0.097	3.307	0.00095	0.035

STIP1	TNFa	29	0.158	3.304	0.00096	0.035
BAZ2B	TGFb	128	-0.051	-3.304	0.00096	0.035
EPHA4	TNFa	61	0.110	3.303	0.00097	0.035
LNPEP	MAPK	38	0.065	3.294	0.001	0.036
CHD6	NFkB	127	0.096	3.261	0.0011	0.038
EPHA4	NFkB	61	0.138	3.260	0.0011	0.038
FAT1	TNFa	239	0.056	3.262	0.0011	0.038
FBXO11	NFkB	41	0.168	3.272	0.0011	0.038
NF1	TNFa	162	0.068	3.273	0.0011	0.038
NRAS	TNFa	56	-0.113	-3.267	0.0011	0.038
PIK3CA	MAPK	667	-0.017	-3.277	0.0011	0.038

Table B2: Reactome vs. mutations (pan-cancer)

Mutation	Pathway	Size	Effect	Wald stat.	P-value	FDR
NFE2L2	p53	105	0.196	7.941	2.6e-15	1.3e-11
BRAF	MAPK	313	0.097	7.525	6.6e-14	1.7e-10
BRAF	JAK-STAT	313	0.098	7.281	4e-13	6.9e-10
KEAP1	p53	114	0.165	6.947	4.4e-12	5.6e-09
NRAS	JAK-STAT	56	-0.210	-6.857	8.2e-12	8.2e-09
TP53	TGFb	1339	0.046	6.810	1.1e-11	8.2e-09
CDH1	TGFb	170	-0.106	-6.810	1.1e-11	8.2e-09
BRAF	Trail	313	0.153	6.543	6.9e-11	4.4e-08
BRAF	NFkB	313	0.089	6.381	2e-10	1.1e-07
PTEN	MAPK	284	-0.084	-6.209	5.9e-10	3e-07
TP53	PI3K	1339	-0.031	-6.172	7.5e-10	3.5e-07
CTNNB1	TNFa	182	-0.112	-5.871	4.7e-09	2e-06
CASP8	JAK-STAT	88	0.140	5.683	1.4e-08	5.6e-06
CTNNB1	MAPK	182	-0.093	-5.615	2.1e-08	7.7e-06
BRAF	VEGF	313	0.065	5.556	2.9e-08	1e-05
GATA3	p53	136	0.120	5.485	4.4e-08	1.4e-05
STK11	Hypoxia	59	0.155	5.421	6.3e-08	1.9e-05
CDH1	p53	170	-0.106	-5.381	7.9e-08	2.2e-05
PIK3CA	MAPK	667	-0.049	-5.265	1.5e-07	4e-05
NRAS	NFkB	56	-0.164	-5.127	3.1e-07	7.5e-05
VHL	p53	209	-0.092	-5.130	3.1e-07	7.5e-05
PTEN	TGFb	284	-0.061	-5.005	5.9e-07	0.00014
MLL2	p53	278	0.078	4.968	7.1e-07	0.00016
PTEN	PI3K	284	-0.045	-4.948	7.8e-07	0.00017
BRAF	PI3K	313	0.041	4.758	2e-06	0.00041
BRAF	EGFR	313	0.044	4.735	2.3e-06	0.00045
STK11	p53	59	0.153	4.638	3.6e-06	0.00069
BRAF	TGFb	313	0.054	4.604	4.3e-06	0.00078
ZFHX3	p53	166	0.091	4.550	5.5e-06	0.00097
CTNNB1	VEGF	182	-0.068	-4.505	6.8e-06	0.0011
NRAS	Trail	56	-0.241	-4.506	6.8e-06	0.0011
EP300	JAK-STAT	123	0.092	4.385	1.2e-05	0.0019
TAF1	p53	111	0.105	4.325	1.6e-05	0.0024
BRAF	TNFa	313	0.064	4.301	1.7e-05	0.0026
CDH1	Hypoxia	170	-0.072	-4.237	2.3e-05	0.0034
ASH1L	TGFb	127	0.076	4.233	2.4e-05	0.0034
TSC1	p53	57	0.137	4.098	4.3e-05	0.0059
NRAS	PI3K	56	-0.081	-4.092	4.4e-05	0.0059
VHL	Trail	209	0.114	4.048	5.3e-05	0.0068
CNOT1	p53	116	0.096	4.044	5.4e-05	0.0068
ASH1L	Hypoxia	127	0.079	4.025	5.8e-05	0.0072
CNOT4	p53	40	0.159	3.990	6.7e-05	0.0079
EP300	NFkB	123	0.087	3.992	6.7e-05	0.0079
BCL11A	Trail	68	0.194	3.986	6.9e-05	0.0079
TP53	EGFR	1339	-0.021	-3.962	7.6e-05	0.0086
BCL11A	p53	68	0.121	3.922	8.9e-05	0.0099
CASP8	Trail	88	0.168	3.909	9.4e-05	0.01
KRAS	Hypoxia	309	0.050	3.874	0.00011	0.012
NR4A2	TGFb	47	0.112	3.864	0.00011	0.012
MMP2	p53	42	0.150	3.843	0.00012	0.013
ASH1L	p53	127	0.087	3.834	0.00013	0.013
PBRM1	NFkB	206	0.065	3.808	0.00014	0.013
PGR	p53	59	0.126	3.814	0.00014	0.013
B2M	JAK-STAT	34	0.150	3.817	0.00014	0.013
CASP8	VEGF	88	0.081	3.795	0.00015	0.014
NRAS	TNFa	56	-0.128	-3.786	0.00016	0.014
TNPO1	p53	42	0.147	3.781	0.00016	0.014
FBXW7	p53	134	0.083	3.759	0.00017	0.015

ERCC2	TGFb	41	0.117	3.742	0.00019	0.016
ARFGEF1	p53	90	0.100	3.727	0.0002	0.017
ANK3	TGFb	174	0.057	3.695	0.00022	0.018
MLL2	TGFb	278	0.046	3.701	0.00022	0.018
CASP8	NFkB	88	0.095	3.693	0.00022	0.018
MACF1	p53	257	0.060	3.679	0.00024	0.019
PIK3CA	TGFb	667	-0.031	-3.677	0.00024	0.019
WHSC1	p53	69	0.111	3.627	0.00029	0.022
NRAS	EGFR	56	-0.076	-3.620	0.0003	0.022
PIK3R1	MAPK	126	-0.072	-3.620	0.0003	0.022
CTCF	MAPK	105	-0.078	-3.605	0.00032	0.023
CDH1	MAPK	170	-0.062	-3.585	0.00034	0.025
BAZ2B	p53	128	0.081	3.572	0.00036	0.026
EPHA2	p53	60	0.116	3.547	0.00039	0.028
MLL	p53	126	0.081	3.550	0.00039	0.028
NR4A2	p53	47	0.131	3.543	0.0004	0.028
NCOR2	p53	95	0.092	3.532	0.00042	0.028
PLXNA1	p53	95	0.092	3.510	0.00045	0.03
PBRM1	Trail	206	0.100	3.503	0.00047	0.031
VHL	Hypoxia	209	0.054	3.495	0.00048	0.031
MGA	p53	131	0.078	3.491	0.00049	0.031
TP53BP1	JAK-STAT	87	0.086	3.473	0.00052	0.033
HLA-A	JAK-STAT	54	0.109	3.470	0.00053	0.033
NOTCH1	JAK-STAT	141	0.068	3.460	0.00055	0.034
SMAD4	Trail	108	0.134	3.450	0.00057	0.035
APC	EGFR	154	-0.044	-3.439	0.00059	0.036
SHMT1	JAK-STAT	19	0.180	3.437	0.0006	0.036
VHL	NFkB	209	0.058	3.434	0.0006	0.036
SHMT1	Trail	19	0.312	3.408	0.00066	0.039
KEAP1	NFkB	114	-0.077	-3.402	0.00068	0.039
EP300	Trail	123	0.124	3.392	0.0007	0.04
NF1	p53	162	0.068	3.390	0.00071	0.04
CBFB	Hypoxia	32	0.131	3.378	0.00074	0.041
PTEN	EGFR	284	-0.032	-3.369	0.00076	0.042
APC	PI3K	154	-0.040	-3.330	0.00088	0.048
EIF1AX	Trail	18	-0.312	-3.319	0.00091	0.049

Table B3: SPIA vs. mutations (pan-cancer)

Mutation	Pathway	Size	Effect	Wald stat.	P-value	FDR
BRAF	EGFR	313	46.908	22.263	3.7e-103	1.4e-99
VHL	NFkB	209	46.685	20.232	2e-86	3.7e-83
PIK3CA	JAK-STAT	667	-3.238	-17.593	1.3e-66	1.6e-63
KRAS	EGFR	309	36.387	16.722	1.4e-60	1.3e-57
VHL	VEGF	209	14.834	16.365	3.5e-58	2.6e-55
BRAF	JAK-STAT	313	4.177	16.338	5.4e-58	3.3e-55
VHL	Trail	209	28.312	14.988	2.4e-49	1.3e-46
VHL	PI3K	209	-4.569	-14.112	4.3e-44	2e-41
TP53	PI3K	1339	2.146	13.790	3.1e-42	1.3e-39
PTEN	JAK-STAT	284	-3.656	-13.532	9.2e-41	3.4e-38
PBRM1	NFkB	206	31.457	13.146	1.3e-38	4.4e-36
VHL	MAPK	209	26.629	13.121	1.8e-38	5.5e-36
KRAS	NFkB	309	-25.626	-12.912	2.4e-37	6.9e-35
VHL	JAK-STAT	209	3.672	11.718	3.6e-31	9.6e-29
PBRM1	VEGF	206	10.480	11.279	4.9e-29	1.2e-26
PBRM1	Trail	206	21.485	11.152	2e-28	4.6e-26
KRAS	VEGF	309	-8.541	-11.085	4.1e-28	8.9e-26
PIK3CA	EGFR	667	-17.154	-10.728	1.8e-26	3.8e-24
KRAS	MAPK	309	-17.813	-10.432	3.9e-25	7.7e-23
PIK3R1	JAK-STAT	126	-4.135	-10.324	1.2e-24	2.2e-22
CDH1	JAK-STAT	170	-3.439	-9.900	7.9e-23	1.4e-20
PTEN	VEGF	284	-7.494	-9.314	2.1e-20	3.5e-18
MAP3K1	JAK-STAT	129	-3.443	-8.658	7.1e-18	1.1e-15
PBRM1	MAPK	206	17.886	8.643	8e-18	1.2e-15
STK11	MAPK	59	-31.427	-8.289	1.6e-16	2.3e-14
TP53	NFkB	1339	-9.533	-8.243	2.3e-16	3.3e-14
KEAP1	MAPK	114	-22.639	-8.236	2.4e-16	3.3e-14
GATA3	MAPK	136	-19.566	-7.743	1.2e-14	1.6e-12
CTNNB1	VEGF	182	-7.644	-7.690	1.9e-14	2.4e-12
NOTCH1	PI3K	140	3.057	7.662	2.3e-14	2.9e-12
CTNNB1	TGFb	182	8.891	7.566	4.8e-14	5.8e-12
BRAF	MAPK	313	12.371	7.234	5.7e-13	6.6e-11
PTEN	TGFb	284	-6.817	-7.134	1.2e-12	1.3e-10

PBRM1	JAK-STAT	206	2.257	7.073	1.8e-12	2e-10
PIK3R1	VEGF	126	-8.367	-7.050	2.1e-12	2.3e-10
CDKN2A	PI3K	165	2.586	7.004	2.9e-12	3e-10
PIK3CA	VEGF	667	-3.912	-6.988	3.3e-12	3.3e-10
FBXW7	JAK-STAT	134	-2.725	-6.956	4.1e-12	4e-10
CDH1	EGFR	170	-20.242	-6.831	9.8e-12	9.4e-10
PBRM1	PI3K	206	-2.267	-6.819	1.1e-11	9.9e-10
NOTCH1	Trail	141	15.746	6.754	1.7e-11	1.5e-09
ARHGAP35	JAK-STAT	59	-3.878	-6.632	3.8e-11	3.3e-09
CASP8	Trail	88	19.248	6.568	5.8e-11	5e-09
BRAF	Trail	313	10.300	6.418	1.6e-10	1.3e-08
GATA3	JAK-STAT	136	-2.409	-6.185	6.9e-10	5.7e-08
KEAP1	NFkB	114	-19.973	-6.180	7.1e-10	5.7e-08
RPL22	JAK-STAT	46	-4.074	-6.158	8.2e-10	6.4e-08
PIK3R1	TGFb	126	-8.608	-6.127	9.9e-10	7.7e-08
ARID1A	JAK-STAT	318	-1.591	-6.083	1.3e-09	9.6e-08
RUNX1	JAK-STAT	49	-3.904	-6.088	1.3e-09	9.6e-08
TP53	JAK-STAT	1339	-0.898	-5.889	4.2e-09	3.1e-07
MLL3	MAPK	296	-10.131	-5.760	9.1e-09	6.5e-07
TGFBR2	PI3K	65	3.328	5.724	1.1e-08	7.9e-07
SMAD4	EGFR	108	21.011	5.690	1.4e-08	9.4e-07
FAT1	JAK-STAT	239	-1.693	-5.672	1.5e-08	1e-06
PTEN	PI3K	284	-1.622	-5.651	1.7e-08	1.1e-06
EP300	Trail	123	14.043	5.629	1.9e-08	1.3e-06
TJP2	TGFb	52	-12.134	-5.601	2.3e-08	1.5e-06
STK11	VEGF	59	-9.578	-5.560	2.9e-08	1.8e-06
FGFR2	JAK-STAT	76	-2.865	-5.539	3.2e-08	2e-06
KRAS	Trail	309	-8.810	-5.449	5.4e-08	3.3e-06
CHD4	JAK-STAT	129	-2.164	-5.410	6.7e-08	4e-06
CASP8	NFkB	88	19.847	5.409	6.8e-08	4e-06
ARID1A	TGFb	318	-4.757	-5.226	1.8e-07	1.1e-05
CTCF	JAK-STAT	105	-2.306	-5.217	1.9e-07	1.1e-05
BAP1	NFkB	89	18.975	5.198	2.1e-07	1.2e-05
EPHA2	Trail	60	18.169	5.128	3.1e-07	1.7e-05
SMAD4	PI3K	108	2.319	5.107	3.4e-07	1.9e-05
FOXA2	JAK-STAT	19	-5.240	-5.102	3.5e-07	1.9e-05
TAF1	JAK-STAT	111	-2.194	-5.097	3.6e-07	1.9e-05
EPHA2	TGFb	60	-10.246	-5.070	4.2e-07	2.2e-05
MLL2	TGFb	278	-4.873	-5.033	5.1e-07	2.6e-05
SPOP	VEGF	79	-7.503	-5.022	5.3e-07	2.7e-05
BCOR	JAK-STAT	97	-2.290	-4.983	6.5e-07	3.3e-05
ZFHX3	JAK-STAT	166	-1.758	-4.957	7.5e-07	3.7e-05
MLL2	Trail	278	8.314	4.896	1e-06	5e-05
CASP8	PI3K	88	2.456	4.894	1e-06	5e-05
NSD1	JAK-STAT	128	-1.955	-4.866	1.2e-06	5.6e-05
UPF3B	JAK-STAT	41	-3.410	-4.860	1.2e-06	5.7e-05
BCLAF1	MAPK	85	-15.459	-4.847	1.3e-06	6.1e-05
FN1	JAK-STAT	121	-1.992	-4.823	1.5e-06	6.7e-05
MYH10	JAK-STAT	107	-2.112	-4.819	1.5e-06	6.8e-05
MKL1	TGFb	42	-11.579	-4.804	1.6e-06	7.1e-05
SIN3A	JAK-STAT	66	-2.668	-4.809	1.6e-06	7.1e-05
NCOR2	TGFb	95	-7.713	-4.778	1.8e-06	7.9e-05
BRAF	PI3K	313	1.316	4.787	1.8e-06	7.7e-05
PTCH1	JAK-STAT	99	-2.156	-4.737	2.3e-06	9.6e-05
STK11	NFkB	59	-21.113	-4.725	2.4e-06	9.9e-05
BAP1	VEGF	89	6.667	4.728	2.4e-06	9.9e-05
PRPF8	TGFb	85	-8.036	-4.715	2.5e-06	0.0001
MYH10	TGFb	107	-7.145	-4.689	2.8e-06	0.00012
ARID5B	JAK-STAT	75	-2.440	-4.680	3e-06	0.00012
B2M	EGFR	34	-30.500	-4.675	3e-06	0.00012
TP53	TGFb	1339	-2.481	-4.673	3.1e-06	0.00012
MTOR	TGFb	136	-6.324	-4.660	3.3e-06	0.00013
SETD2	NFkB	161	12.777	4.657	3.3e-06	0.00013
CDKN2A	TGFb	165	-5.731	-4.632	3.7e-06	0.00014
MAP2K4	JAK-STAT	64	-2.599	-4.612	4.1e-06	0.00015
TGFBR2	Trail	65	15.729	4.614	4.1e-06	0.00015
CIC	TGFb	74	-8.375	-4.591	4.6e-06	0.00017
ASPM	JAK-STAT	178	-1.574	-4.585	4.7e-06	0.00017
FGFR3	MAPK	46	-19.696	-4.566	5.1e-06	0.00019
MLH3	JAK-STAT	59	-2.676	-4.562	5.2e-06	0.00019
CREBBP	PI3K	145	1.793	4.547	5.6e-06	0.0002
STAG2	TGFb	92	-7.445	-4.539	5.8e-06	0.00021
ZFHX3	TGFb	166	-5.539	-4.489	7.4e-06	0.00026
FAM123B	JAK-STAT	78	-2.294	-4.485	7.5e-06	0.00026
MLL3	EGFR	296	-10.250	-4.467	8.2e-06	0.00028

ANK3	JAK-STAT	174	-1.533	-4.418	1e-05	0.00035
ASH1L	TGFb	127	-6.175	-4.401	1.1e-05	0.00037
EPHA4	TGFb	61	-8.777	-4.375	1.2e-05	0.0004
KALRN	JAK-STAT	152	-1.623	-4.385	1.2e-05	0.00039
KRAS	TGFb	309	-4.055	-4.392	1.2e-05	0.00038
MYH9	JAK-STAT	110	-1.898	-4.387	1.2e-05	0.00039
NSD1	TGFb	128	-6.140	-4.392	1.2e-05	0.00038
AKAP9	Trail	169	9.369	4.367	1.3e-05	0.00041
ANK3	Trail	174	9.247	4.370	1.3e-05	0.00041
FAT1	PI3K	238	1.353	4.337	1.5e-05	0.00047
CASP8	JAK-STAT	88	-2.090	-4.334	1.5e-05	0.00047
PTEN	NFkB	284	9.086	4.320	1.6e-05	0.0005
CTCF	VEGF	105	-5.619	-4.317	1.6e-05	0.0005
DHX35	TGFb	38	-10.832	-4.275	2e-05	0.0006
STAG2	MAPK	92	-13.084	-4.261	2.1e-05	0.00063
ZNF750	PI3K	44	3.006	4.257	2.1e-05	0.00063
SMAD4	TGFb	108	-6.444	-4.246	2.2e-05	0.00066
GATA3	EGFR	136	-14.046	-4.244	2.3e-05	0.00066
VHL	EGFR	209	-11.426	-4.235	2.3e-05	0.00068
FGFR1	Trail	34	19.759	4.208	2.6e-05	0.00076
BRWD1	JAK-STAT	105	-1.864	-4.210	2.6e-05	0.00076
FOXA1	JAK-STAT	56	-2.531	-4.205	2.7e-05	0.00076
ATM	JAK-STAT	189	-1.404	-4.207	2.7e-05	0.00076
CUX1	TGFb	110	-6.291	-4.182	3e-05	0.00083
IREB2	TGFb	52	-9.044	-4.166	3.2e-05	0.00088
KEAP1	VEGF	114	-5.166	-4.129	3.7e-05	0.001
ARFGEF2	JAK-STAT	85	-2.024	-4.125	3.8e-05	0.001
NRAS	EGFR	56	20.999	4.116	3.9e-05	0.0011
PIK3R1	NFkB	126	12.655	4.098	4.3e-05	0.0012
PER1	TGFb	46	-9.425	-4.087	4.5e-05	0.0012
NOTCH1	JAK-STAT	141	-1.563	-4.070	4.8e-05	0.0013
MAP4K1	MAPK	38	-19.237	-4.055	5.1e-05	0.0014
PGR	TGFb	59	-8.254	-4.046	5.3e-05	0.0014
LAMA2	JAK-STAT	173	-1.403	-4.028	5.7e-05	0.0015
CRNKL1	JAK-STAT	51	-2.537	-4.024	5.8e-05	0.0015
FBXW7	TGFb	134	-5.503	-4.023	5.9e-05	0.0015
BAZ2B	TGFb	128	-5.623	-4.021	5.9e-05	0.0015
EZH2	TGFb	37	-10.312	-4.015	6.1e-05	0.0015
ARFGEF1	JAK-STAT	90	-1.916	-4.015	6.1e-05	0.0015
CHD3	JAK-STAT	107	-1.758	-4.007	6.3e-05	0.0016
KDM5C	Trail	78	12.438	3.987	6.8e-05	0.0017
FAM46C	TGFb	15	-16.029	-3.985	6.9e-05	0.0017

Table B4: Pathifier vs. mutations (pan-cancer)

Mutation	Pathway	Size	Effect	Wald stat.	P-value	FDR
PTEN	MAPK	284	0.177	17.422	2.1e-65	1.1e-61
PTEN	TGFb	284	0.180	16.954	3.7e-62	9.5e-59
PTEN	VEGF	284	0.160	15.553	6.9e-53	1.2e-49
PTEN	NFkB	284	0.190	15.125	3.4e-50	4.3e-47
VHL	JAK-STAT	209	0.190	14.768	5.3e-48	5.4e-45
VHL	p53	209	0.220	14.697	1.4e-47	1.2e-44
PTEN	TNFa	284	0.158	14.223	9.7e-45	7e-42
PTEN	JAK-STAT	284	0.154	13.707	9.4e-42	6e-39
VHL	PI3K	209	0.171	12.862	4.5e-37	2.6e-34
VHL	NFkB	209	0.185	12.610	9.9e-36	5e-33
CTNNB1	TNFa	182	0.160	11.634	9.4e-31	4.4e-28
PIK3R1	TGFb	126	0.183	11.521	3.3e-30	1.4e-27
PIK3R1	MAPK	126	0.174	11.390	1.5e-29	5.7e-27
PIK3R1	VEGF	126	0.171	11.152	2e-28	7.2e-26
PBRM1	p53	206	0.166	10.890	3.3e-27	1.1e-24
ARID1A	JAK-STAT	318	0.112	10.427	4.1e-25	1.3e-22
CTNNB1	TGFb	182	0.139	10.416	4.6e-25	1.4e-22
TP53	EGFR	1339	0.054	10.226	3.2e-24	8.9e-22
CTNNB1	PI3K	182	0.146	10.218	3.4e-24	9.2e-22
PTEN	p53	284	0.134	10.192	4.4e-24	1.1e-21
PIK3R1	NFkB	126	0.191	10.185	4.8e-24	1.2e-21
PIK3R1	TNFa	126	0.168	10.169	5.6e-24	1.3e-21
PBRM1	JAK-STAT	206	0.128	9.723	4.4e-22	9.3e-20
ARID1A	TGFb	318	0.100	9.723	4.4e-22	9.3e-20
PIK3R1	JAK-STAT	126	0.155	9.262	3.3e-20	6.8e-18
ARID1A	VEGF	318	0.091	9.148	9.3e-20	1.8e-17
CTNNB1	EGFR	182	0.107	9.145	9.6e-20	1.8e-17

VHL	Hypoxia	209	0.112	9.090	1.6e-19	2.9e-17
VHL	Trail	209	0.137	8.938	6.1e-19	1.1e-16
PIK3CA	Trail	667	0.080	8.671	6.3e-18	1.1e-15
CTCF	TGFb	105	0.151	8.637	8.4e-18	1.4e-15
ARID1A	p53	318	0.106	8.449	4.2e-17	6.6e-15
NOTCH1	p53	141	0.152	8.231	2.5e-16	3.9e-14
KRAS	TNFa	309	0.088	8.137	5.5e-16	8.2e-14
TP53	JAK-STAT	1339	-0.051	-8.031	1.3e-15	1.9e-13
MLL2	p53	278	0.107	7.989	1.8e-15	2.5e-13
PIK3CA	VEGF	667	0.058	7.880	4.3e-15	5.9e-13
PBRM1	NFkB	206	0.117	7.858	5.1e-15	6.8e-13
FAT1	PI3K	239	-0.097	-7.634	2.9e-14	3.8e-12
VHL	TNFa	209	0.097	7.461	1.1e-13	1.4e-11
PTEN	Trail	284	0.099	7.412	1.5e-13	1.9e-11
PTEN	EGFR	284	0.071	7.410	1.6e-13	1.9e-11
BCOR	TGFb	97	0.134	7.371	2.1e-13	2.5e-11
ARHGAP35	MAPK	59	0.164	7.331	2.8e-13	3.2e-11
ZFHX3	TGFb	166	0.103	7.313	3.2e-13	3.6e-11
TP53	PI3K	1339	-0.047	-7.293	3.7e-13	4.1e-11
KRAS	p53	309	-0.093	-7.250	5.1e-13	5.5e-11
NFE2L2	MAPK	105	0.122	7.238	5.5e-13	5.9e-11
MLL2	PI3K	278	-0.086	-7.226	6e-13	6.3e-11
RPL22	NFkB	46	0.222	7.183	8.2e-13	8.4e-11
ARHGAP35	NFkB	59	0.194	7.097	1.5e-12	1.5e-10
RPL22	MAPK	46	0.179	7.095	1.5e-12	1.5e-10
PIK3R1	p53	126	0.137	7.052	2.1e-12	2e-10
ARHGAP35	TGFb	59	0.163	7.030	2.4e-12	2.3e-10
APC	TGFb	154	0.102	7.020	2.6e-12	2.4e-10
CTCF	VEGF	105	0.118	6.974	3.6e-12	3.3e-10
KRAS	JAK-STAT	309	-0.077	-6.958	4.1e-12	3.6e-10
CASP8	PI3K	88	-0.143	-6.957	4.1e-12	3.6e-10
NOTCH1	PI3K	141	-0.112	-6.878	7.1e-12	6.1e-10
RPL22	TGFb	46	0.181	6.877	7.1e-12	6.1e-10
BAP1	JAK-STAT	89	0.137	6.874	7.3e-12	6.1e-10
CTNNB1	NFkB	182	0.109	6.863	7.9e-12	6.5e-10
RPL22	VEGF	46	0.174	6.861	8e-12	6.5e-10
ZFHX3	VEGF	166	0.093	6.820	1.1e-11	8.5e-10
TP53	VEGF	1339	0.040	6.792	1.3e-11	1e-09
CTCF	p53	105	0.144	6.760	1.6e-11	1.2e-09
CTNNB1	p53	182	0.110	6.733	1.9e-11	1.5e-09
ARHGAP35	TNFa	59	0.162	6.728	2e-11	1.5e-09
CDKN2A	PI3K	165	-0.101	-6.668	3e-11	2.2e-09
PBRM1	Hypoxia	206	0.082	6.585	5.2e-11	3.8e-09
PPP2R1A	TGFb	67	0.143	6.566	5.9e-11	4.2e-09
CTCF	JAK-STAT	105	0.120	6.548	6.6e-11	4.7e-09
FBXW7	VEGF	134	0.099	6.542	6.9e-11	4.8e-09
FAT1	p53	239	0.094	6.532	7.4e-11	5.1e-09
ARID5B	TGFb	75	0.135	6.524	7.8e-11	5.3e-09
EP300	JAK-STAT	123	0.111	6.515	8.2e-11	5.5e-09
CTCF	TNFa	105	0.118	6.512	8.4e-11	5.5e-09
BRAF	VEGF	313	0.066	6.511	8.5e-11	5.5e-09
RPL22	TNFa	46	0.177	6.504	8.9e-11	5.7e-09
PIK3CA	MAPK	667	0.047	6.485	1e-10	6.4e-09
CEP290	VEGF	88	0.119	6.410	1.6e-10	1e-08
BCOR	JAK-STAT	97	0.122	6.402	1.7e-10	1.1e-08
FOXA2	VEGF	19	0.251	6.373	2.1e-10	1.3e-08
CUX1	TGFb	110	0.109	6.356	2.3e-10	1.4e-08
ARHGAP35	VEGF	59	0.142	6.296	3.4e-10	2.1e-08
VHL	VEGF	209	-0.077	-6.278	3.8e-10	2.3e-08
SETD2	JAK-STAT	161	0.093	6.239	4.9e-10	2.9e-08
KEAP1	Trail	114	-0.128	-6.230	5.2e-10	3e-08
MTOR	TGFb	136	0.096	6.218	5.6e-10	3.2e-08
PIK3CA	JAK-STAT	667	0.049	6.168	7.6e-10	4.3e-08
ARHGAP35	JAK-STAT	59	0.150	6.151	8.5e-10	4.8e-08
ARID5B	JAK-STAT	75	0.133	6.137	9.3e-10	5.2e-08
TP53	Trail	1339	-0.045	-6.134	9.5e-10	5.2e-08
FGFR2	VEGF	76	0.122	6.127	9.9e-10	5.4e-08
MLL	p53	126	0.119	6.090	1.2e-09	6.7e-08
BAP1	p53	89	0.140	6.067	1.4e-09	7.6e-08
FOXA2	TGFb	19	0.247	6.059	1.5e-09	7.9e-08
SMARCA4	p53	118	0.122	6.056	1.5e-09	8e-08
FBXW7	TGFb	134	0.094	6.046	1.6e-09	8.4e-08
MLL2	TGFb	278	0.067	6.019	1.9e-09	9.8e-08
FGFR2	TNFa	76	0.128	6.017	2e-09	9.9e-08
CTCF	MAPK	105	0.101	6.004	2.1e-09	1.1e-07

BRWD1	TGFb	105	0.105	5.990	2.3e-09	1.1e-07
RPL22	JAK-STAT	46	0.164	5.962	2.7e-09	1.3e-07
ARID1A	TNFa	318	0.064	5.939	3.1e-09	1.5e-07
TAF1	TGFb	111	0.101	5.933	3.2e-09	1.6e-07
CTNNB1	MAPK	182	0.077	5.931	3.3e-09	1.6e-07
EGFR	Trail	99	-0.130	-5.929	3.3e-09	1.6e-07
MTOR	p53	136	0.111	5.920	3.5e-09	1.6e-07
KEAP1	p53	114	0.121	5.889	4.2e-09	1.9e-07
ARID1A	PI3K	318	-0.066	-5.890	4.2e-09	1.9e-07
BRAF	p53	313	-0.075	-5.888	4.3e-09	1.9e-07
EPHA2	PI3K	60	-0.146	-5.884	4.4e-09	2e-07
NSD1	PI3K	128	-0.101	-5.881	4.4e-09	2e-07
VHL	MAPK	209	0.071	5.875	4.6e-09	2e-07
FGFR2	TGFb	76	0.121	5.872	4.7e-09	2.1e-07
PTEN	Hypoxia	284	0.063	5.871	4.7e-09	2.1e-07
CUX1	p53	110	0.122	5.864	4.9e-09	2.1e-07
PIK3CA	TGFb	667	0.044	5.845	5.5e-09	2.4e-07
BRWD1	VEGF	105	0.099	5.833	5.9e-09	2.5e-07
CUX1	JAK-STAT	110	0.104	5.815	6.6e-09	2.8e-07
BRAF	TNFa	313	0.063	5.813	6.6e-09	2.8e-07
PPP2R1A	MAPK	67	0.122	5.797	7.3e-09	3e-07
BAZ2B	TGFb	128	0.093	5.795	7.4e-09	3.1e-07
PBRM1	Trail	206	0.090	5.785	7.9e-09	3.2e-07
CEP290	TGFb	88	0.110	5.769	8.6e-09	3.5e-07
TP53	MAPK	1339	0.034	5.754	9.4e-09	3.8e-07
ARID5B	p53	75	0.144	5.748	9.7e-09	3.9e-07
FAT1	TGFb	239	0.068	5.724	1.1e-08	4.4e-07
FN1	TGFb	121	0.094	5.712	1.2e-08	4.7e-07
ARID1A	Hypoxia	318	0.058	5.712	1.2e-08	4.7e-07
CASP8	p53	88	0.133	5.702	1.3e-08	4.9e-07
MACF1	TGFb	257	0.065	5.683	1.4e-08	5.5e-07
LAMA2	TGFb	173	0.078	5.680	1.5e-08	5.5e-07
SETD2	p53	161	0.099	5.677	1.5e-08	5.6e-07
BCOR	VEGF	97	0.100	5.673	1.5e-08	5.7e-07
CLSPN	MAPK	43	0.148	5.668	1.6e-08	5.8e-07
CYLD	TGFb	61	0.129	5.649	1.7e-08	6.4e-07
TAF1	MAPK	111	0.093	5.651	1.7e-08	6.4e-07
STK11	Trail	59	-0.160	-5.632	1.9e-08	7e-07
RASA1	p53	87	0.131	5.609	2.2e-08	7.9e-07
SETD2	NFkB	161	0.094	5.597	2.3e-08	8.4e-07
CSDE1	TGFb	51	0.140	5.593	2.4e-08	8.5e-07
ANK3	VEGF	174	0.074	5.583	2.5e-08	8.9e-07
SUZ12	NFkB	35	0.198	5.583	2.5e-08	8.9e-07
VHL	TGFb	209	0.070	5.566	2.8e-08	9.8e-07
MED17	TGFb	29	0.184	5.556	3e-08	1e-06
FGFR2	JAK-STAT	76	0.119	5.525	3.5e-08	1.2e-06
PPP2R1A	TNFa	67	0.125	5.515	3.7e-08	1.3e-06
TAF1	VEGF	111	0.091	5.504	4e-08	1.3e-06

Table B5: PARADIGM vs. mutations (pan-cancer)

Mutation	Pathway	Size	Effect	Wald stat.	P-value	FDR
PTEN	PI3K	281	-0.973	-8.811	1.9e-18	8.8e-15
TP53	PI3K	1337	0.461	7.509	7.5e-14	1.7e-10
PTEN	TGFb	281	-0.582	-7.126	1.2e-12	1.9e-09
PIK3R1	PI3K	124	-1.080	-6.618	4.2e-11	4.8e-08
PTEN	MAPK	281	-0.364	-6.376	2.1e-10	1.9e-07
BRAF	MAPK	312	0.340	6.241	4.8e-10	3.7e-07
CTNNB1	TGFb	182	-0.574	-5.721	1.1e-08	7.6e-06
PIK3CA	MAPK	663	-0.212	-5.354	9.1e-08	5.3e-05
CTCF	MAPK	104	-0.485	-5.298	1.2e-07	6.4e-05
GATA3	MAPK	136	-0.419	-5.209	2e-07	8.9e-05
ZFHX3	TGFb	166	-0.545	-5.200	2.1e-07	8.9e-05
CTNNB1	PI3K	182	-0.701	-5.150	2.7e-07	0.00011
TP53	MAPK	1337	0.161	5.072	4.1e-07	0.00015
PTEN	JAK-STAT	281	-0.379	-4.860	1.2e-06	0.00041
PTEN	p53	281	-0.215	-4.800	1.7e-06	0.00051
CTCF	PI3K	104	-0.814	-4.568	5.1e-06	0.0015
HDAC3	VEGF	22	0.045	4.506	6.8e-06	0.0019
ARID1A	TGFb	316	-0.340	-4.375	1.2e-05	0.003
CTNNB1	MAPK	182	-0.306	-4.369	1.3e-05	0.003
BRAF	PI3K	312	0.463	4.364	1.3e-05	0.003
CTCF	TGFb	104	-0.541	-4.118	3.9e-05	0.0086

ARID1A	MAPK	316	-0.222	-4.094	4.3e-05	0.0091
APC	TGFb	151	-0.446	-4.058	5.1e-05	0.01
BRAF	JAK-STAT	312	0.299	4.016	6e-05	0.012
PIK3R1	TGFb	124	-0.480	-3.978	7.1e-05	0.013
GATA3	TGFb	136	-0.453	-3.920	9e-05	0.016
PIK3CA	p53	663	-0.121	-3.879	0.00011	0.018
ZFHX3	MAPK	166	-0.283	-3.869	0.00011	0.018
NRAS	JAK-STAT	56	-0.653	-3.858	0.00012	0.018
ZNF292	MAPK	109	-0.346	-3.857	0.00012	0.018
PTEN	Hypoxia	281	-0.301	-3.661	0.00025	0.036
CDH1	PI3K	170	-0.517	-3.669	0.00025	0.036
CDH1	TGFb	170	0.380	3.663	0.00025	0.036
SETDB1	TGFb	68	-0.590	-3.644	0.00027	0.037
CTNNB1	JAK-STAT	182	-0.347	-3.627	0.00029	0.039
ASH1L	MAPK	127	-0.301	-3.615	0.0003	0.039
CDH1	EGFR	170	0.025	3.595	0.00033	0.041
DIS3	VEGF	41	0.026	3.574	0.00036	0.043
NUP93	VEGF	41	0.026	3.539	0.00041	0.048

Table B6: Perturbation-response genes vs. mutations (pan-cancer)

Mutation	Pathway	Size	Effect	Wald stat.	P-value	FDR
VHL	Hypoxia	209	2.201	33.612	1.8e-216	9.3e-213
TP53	PI3K	1339	0.778	26.882	1.4e-145	3.5e-142
PTEN	p53	284	-1.121	-21.776	4.9e-99	8.3e-96
TP53	EGFR	1339	0.677	21.746	8.7e-99	1.1e-95
CTNNB1	TGFb	182	-1.283	-20.448	3.8e-88	3.9e-85
PTEN	PI3K	284	1.085	19.968	2.4e-84	2.1e-81
CTNNB1	p53	182	-1.196	-18.578	9.7e-74	7.1e-71
PBRM1	Hypoxia	206	1.329	18.436	1.1e-72	6.9e-70
PIK3R1	p53	126	-1.381	-17.949	3.8e-69	2.1e-66
PTEN	TGFb	284	-0.915	-17.711	1.9e-67	9.7e-65
TP53	TNFa	1339	0.476	17.266	2.6e-64	1.2e-61
CTNNB1	Trail	182	-1.176	-16.494	4.9e-59	2.1e-56
PTEN	MAPK	284	-0.987	-16.414	1.7e-58	6.5e-56
TP53	NFkB	1339	0.502	16.283	1.2e-57	4.5e-55
PTEN	Trail	284	-0.937	-16.170	7e-57	2.4e-54
BRAF	PI3K	313	-0.843	-15.941	2.2e-55	6.9e-53
TP53	MAPK	1339	0.529	15.880	5.4e-55	1.6e-52
KRAS	EGFR	309	0.868	15.585	4.3e-53	1.2e-50
PTEN	JAK-STAT	284	-0.884	-15.431	4.1e-52	1.1e-49
ARID1A	PI3K	318	0.805	15.286	3.4e-51	8.6e-49
PIK3R1	PI3K	126	1.241	15.240	6.4e-51	1.6e-48
CTNNB1	MAPK	182	-1.114	-14.963	3.4e-49	7.8e-47
CTNNB1	Hypoxia	182	-1.138	-14.656	2.5e-47	5.7e-45
CTNNB1	JAK-STAT	182	-0.987	-13.920	5.7e-43	1.2e-40
FAT1	PI3K	239	0.832	13.774	3.9e-42	8e-40
ARHGAP35	p53	59	-1.533	-13.517	1.1e-40	2.2e-38
CDKN2A	EGFR	165	0.996	13.216	5.3e-39	1e-36
MLL2	PI3K	278	0.714	12.641	6.9e-36	1.3e-33
PTEN	TNFa	284	-0.641	-12.637	7.2e-36	1.3e-33
PIK3R1	Trail	126	-1.089	-12.626	8.2e-36	1.4e-33
PIK3R1	TGFb	126	-0.974	-12.602	1.1e-35	1.8e-33
TP53	VEGF	1339	0.390	12.268	6.1e-34	9.7e-32
BRAF	MAPK	313	0.712	12.189	1.5e-33	2.4e-31
PTEN	VEGF	284	0.692	12.054	7.5e-33	1.1e-30
CTNNB1	NFkB	182	-0.829	-11.900	4.5e-32	6.6e-30
FBXW7	PI3K	134	0.951	11.896	4.7e-32	6.7e-30
PTEN	Hypoxia	284	-0.742	-11.656	7.4e-31	1e-28
FAT1	VEGF	239	0.724	11.617	1.1e-30	1.5e-28
RPL22	p53	46	-1.487	-11.520	3.4e-30	4.5e-28
CTNNB1	TNFa	182	-0.720	-11.492	4.7e-30	5.9e-28
CTNNB1	PI3K	182	0.781	11.286	4.5e-29	5.6e-27
NSD1	PI3K	128	0.916	11.184	1.4e-28	1.7e-26
NOTCH1	VEGF	141	0.875	10.917	2.5e-27	2.9e-25
NOTCH1	PI3K	141	0.852	10.895	3.1e-27	3.6e-25
CDKN2A	PI3K	165	0.783	10.788	9.7e-27	1.1e-24
TP53	JAK-STAT	1339	0.347	10.786	9.9e-27	1.1e-24
NOTCH1	EGFR	141	0.882	10.777	1.1e-26	1.2e-24
PTEN	NFkB	284	-0.606	-10.668	3.4e-26	3.7e-24
VHL	PI3K	209	-0.687	-10.584	8.2e-26	8.6e-24
ARHGAP35	PI3K	59	1.262	10.542	1.3e-25	1.3e-23
EP300	PI3K	123	0.880	10.523	1.5e-25	1.5e-23

FAT1	EGFR	239	0.671	10.518	1.6e-25	1.6e-23
VHL	VEGF	209	-0.700	-10.520	1.6e-25	1.6e-23
RPL22	PI3K	46	1.400	10.342	9.9e-25	9.3e-23
KALRN	PI3K	152	0.775	10.248	2.5e-24	2.3e-22
CASP8	PI3K	88	1.003	10.182	4.9e-24	4.5e-22
PIK3R1	MAPK	126	-0.914	-10.145	7.1e-24	6.4e-22
LAMA2	PI3K	173	0.721	10.139	7.6e-24	6.7e-22
CTCF	p53	105	-0.866	-10.012	2.7e-23	2.3e-21
CTNNB1	VEGF	182	0.711	10.001	3e-23	2.5e-21
PIK3R1	JAK-STAT	126	-0.853	-9.968	4.1e-23	3.4e-21
VHL	EGFR	209	-0.675	-9.930	6e-23	4.9e-21
KRAS	PI3K	309	0.539	9.926	6.2e-23	5e-21
PIK3R1	VEGF	126	0.835	9.844	1.4e-22	1.1e-20
ARID1A	Hypoxia	318	-0.599	-9.844	1.4e-22	1.1e-20
ARHGAP35	TGFb	59	-1.106	-9.797	2.2e-22	1.7e-20
ANK3	PI3K	174	0.692	9.742	3.7e-22	2.8e-20
ARID1A	TGFb	318	-0.486	-9.640	9.7e-22	7.3e-20
ARID1A	p53	318	-0.491	-9.567	1.9e-21	1.4e-19
ARID5B	PI3K	75	1.020	9.563	2e-21	1.5e-19
CREBBP	PI3K	145	0.731	9.431	7e-21	5e-19
CTCF	PI3K	105	0.852	9.409	8.5e-21	6e-19
ASPM	PI3K	178	0.660	9.383	1.1e-20	7.6e-19
VHL	Trail	209	0.633	9.250	3.7e-20	2.5e-18
SPOP	MAPK	79	-1.038	-9.151	9.1e-20	6.2e-18
PIK3R1	Hypoxia	126	-0.858	-9.116	1.2e-19	8.4e-18
FAT2	PI3K	188	0.618	9.014	3.1e-19	2.1e-17
ZFHX3	PI3K	166	0.655	9.011	3.2e-19	2.1e-17
SMAD4	EGFR	108	0.841	8.998	3.6e-19	2.3e-17
CTCF	TGFb	105	-0.766	-8.980	4.2e-19	2.7e-17
ARHGAP35	Trail	59	-1.132	-8.977	4.3e-19	2.7e-17
ARID1A	Trail	318	-0.504	-8.938	6.1e-19	3.8e-17
NSD1	VEGF	128	0.751	8.902	8.4e-19	5.2e-17
MLL2	EGFR	278	0.531	8.900	8.6e-19	5.2e-17
RASA1	PI3K	87	0.884	8.895	9e-19	5.4e-17
AKAP9	PI3K	169	0.641	8.882	1e-18	5.9e-17
CDKN2A	TNFa	165	0.576	8.708	4.6e-18	2.7e-16
PIK3R1	TNFa	126	-0.653	-8.682	5.7e-18	3.3e-16
MACF1	PI3K	257	0.514	8.680	5.8e-18	3.4e-16
CDKN2A	MAPK	165	0.686	8.633	8.7e-18	4.9e-16
FGFR2	PI3K	76	0.916	8.621	9.7e-18	5.4e-16
AHNAK	PI3K	195	0.581	8.617	1e-17	5.6e-16
TAF1	PI3K	111	0.759	8.589	1.3e-17	7e-16
PIK3CA	Hypoxia	667	-0.381	-8.574	1.4e-17	7.8e-16
CDKN2A	VEGF	165	0.634	8.478	3.3e-17	1.8e-15
BRAF	EGFR	313	0.478	8.448	4.2e-17	2.2e-15
KRAS	p53	309	-0.437	-8.378	7.5e-17	4e-15
ARID1A	VEGF	318	0.461	8.373	7.9e-17	4.1e-15
CDKN2A	NFkB	165	0.616	8.365	8.4e-17	4.3e-15
RB1	PI3K	125	0.695	8.332	1.1e-16	5.6e-15
SPTAN1	PI3K	111	0.735	8.315	1.3e-16	6.4e-15
MLL2	VEGF	278	0.486	8.301	1.4e-16	7.1e-15
ACACA	PI3K	97	0.779	8.252	2.1e-16	1e-14
NOTCH1	NFkB	141	0.656	8.256	2.1e-16	1e-14
APC	PI3K	154	0.623	8.245	2.3e-16	1.1e-14
RPL22	TGFb	46	-1.049	-8.192	3.5e-16	1.7e-14
PIK3CA	PI3K	667	0.321	8.185	3.7e-16	1.8e-14
ARHGAP35	MAPK	59	-1.072	-8.179	3.9e-16	1.8e-14
PPP2R1A	p53	67	-0.883	-8.164	4.4e-16	2.1e-14
NRAS	TNFa	56	-0.907	-8.103	7.2e-16	3.3e-14
ATRX	PI3K	133	0.656	8.092	7.9e-16	3.6e-14
CEP290	PI3K	88	0.800	8.076	9e-16	4.1e-14
PPP2R1A	PI3K	67	0.909	8.037	1.2e-15	5.6e-14
CSDE1	PI3K	51	1.036	8.005	1.6e-15	7.1e-14
EP300	VEGF	123	0.689	7.995	1.7e-15	7.6e-14
FN1	PI3K	121	0.677	7.977	2e-15	8.7e-14
PIK3CA	MAPK	667	-0.340	-7.955	2.4e-15	1e-13
CHD3	PI3K	107	0.716	7.947	2.5e-15	1.1e-13
CHD4	PI3K	129	0.654	7.946	2.5e-15	1.1e-13
FGFR2	p53	76	-0.808	-7.942	2.6e-15	1.1e-13
SIN3A	PI3K	66	0.900	7.894	3.8e-15	1.6e-13
NCOR2	PI3K	95	0.753	7.888	4e-15	1.7e-13
ATR	PI3K	117	0.679	7.878	4.3e-15	1.8e-13
RPL22	TNFa	46	-0.972	-7.875	4.5e-15	1.8e-13
SMARCA4	PI3K	118	0.672	7.826	6.5e-15	2.7e-13
RPL22	JAK-STAT	46	-1.101	-7.817	7e-15	2.8e-13

BCOR	PI3K	97	0.737	7.803	7.8e-15	3.1e-13
CHD8	PI3K	96	0.735	7.740	1.3e-14	5e-13
LRP6	PI3K	70	0.858	7.740	1.3e-14	5e-13
CIC	PI3K	74	0.832	7.712	1.6e-14	6.2e-13
MYH10	PI3K	107	0.694	7.700	1.7e-14	6.8e-13
DICER1	PI3K	88	0.757	7.639	2.8e-14	1.1e-12
ASH1L	PI3K	127	0.630	7.589	4e-14	1.6e-12
ABCB1	PI3K	111	0.672	7.585	4.2e-14	1.6e-12
PIK3CA	VEGF	667	0.304	7.557	5.2e-14	2e-12
KAT6B	p53	42	-1.024	-7.512	7.3e-14	2.7e-12
PLXNA1	PI3K	95	0.717	7.505	7.7e-14	2.9e-12
ASXL1	PI3K	87	0.745	7.476	9.5e-14	3.5e-12
BAZ2B	PI3K	128	0.618	7.474	9.6e-14	3.5e-12
CUX1	PI3K	110	0.664	7.461	1.1e-13	3.9e-12
HDAC9	PI3K	80	0.775	7.460	1.1e-13	3.9e-12
INPPL1	PI3K	69	0.831	7.446	1.2e-13	4.2e-12
ARHGAP35	JAK-STAT	59	-0.928	-7.448	1.2e-13	4.2e-12
TRIO	PI3K	117	0.642	7.442	1.2e-13	4.3e-12
PTPRF	PI3K	80	0.772	7.430	1.3e-13	4.7e-12
NOTCH1	TNFa	141	0.531	7.428	1.4e-13	4.7e-12
BRCA2	PI3K	139	0.589	7.417	1.5e-13	5.1e-12
RPL22	Trail	46	-1.057	-7.385	1.9e-13	6.5e-12
TAF1	TGFb	111	-0.615	-7.382	1.9e-13	6.5e-12
FBXW7	VEGF	134	0.609	7.357	2.3e-13	7.9e-12

B.2 *Pathway scores and CNAs*

Table B7: Gene Ontology vs. CNAs (pan-cancer)

CNA	Pathway	Size	Effect	Wald stat.	P-value	FDR
CTTN_amp	Trail	564	0.352	17.309	7.9e-66	7.9e-62
CCND1_amp	Trail	616	0.298	15.189	2.5e-51	1.2e-47
MYC_amp	NFkB	974	0.097	8.424	4.6e-17	1.5e-13
CDKN2A_del	Trail	1060	0.132	8.342	8.9e-17	2.2e-13
INPPL1_amp	Trail	201	0.277	8.237	2.1e-16	4.1e-13
CDKN2A_del	NFkB	1060	0.081	7.362	2e-13	3.4e-10
NDRG1_amp	NFkB	744	0.091	7.089	1.5e-12	2.2e-09
PIK3CA_amp	VEGF	735	-0.123	-7.043	2.1e-12	2.6e-09
RAD21_amp	NFkB	722	0.092	7.021	2.5e-12	2.7e-09
MYC_amp	p53	974	0.060	6.759	1.5e-11	1.5e-08
TBL1XR1_amp	VEGF	710	-0.119	-6.708	2.1e-11	1.9e-08
FXR1_amp	VEGF	705	-0.119	-6.693	2.4e-11	2e-08
MSR1_del	Trail	392	-0.164	-6.619	4e-11	3e-08
EIF4G1_amp	VEGF	665	-0.115	-6.346	2.4e-10	1.7e-07
NDRG1_amp	p53	744	0.062	6.252	4.4e-10	2.9e-07
MUC20_amp	VEGF	647	-0.111	-6.043	1.6e-09	9.9e-07
DLG1_amp	VEGF	654	-0.110	-5.989	2.2e-09	1.2e-06
CHD1L_amp	EGFR	456	-0.124	-5.977	2.4e-09	1.2e-06
TXNIP_amp	EGFR	456	-0.124	-5.977	2.4e-09	1.2e-06
ASXL1_amp	VEGF	246	-0.172	-5.978	2.4e-09	1.2e-06
NOTCH2_amp	EGFR	481	-0.119	-5.919	3.4e-09	1.6e-06
CDKN2A_del	TNFa	1060	0.051	5.900	3.8e-09	1.7e-06
ASH1L_amp	PI3K	451	-0.068	-5.730	1.1e-08	4.6e-06
MNDA_amp	EGFR	370	-0.128	-5.592	2.4e-08	9.8e-06
RB1_del	PI3K	368	-0.073	-5.526	3.4e-08	1.4e-05
RFC4_amp	VEGF	637	-0.101	-5.461	4.9e-08	1.9e-05
ERBB2_amp	JAK-STAT	378	-0.081	-5.443	5.4e-08	2e-05
PABPC1_amp	NFkB	664	0.074	5.438	5.6e-08	2e-05
PABPC1_amp	p53	664	0.056	5.434	5.8e-08	2e-05
SETDB1_amp	EGFR	513	-0.107	-5.417	6.3e-08	2.1e-05
ERBB2_amp	Trail	378	0.135	5.395	7.1e-08	2.3e-05
FCRL4_amp	EGFR	368	-0.123	-5.373	8e-08	2.5e-05
PIP5K1A_amp	PI3K	504	-0.061	-5.369	8.3e-08	2.5e-05
FRG1B_amp	VEGF	239	-0.156	-5.358	8.8e-08	2.6e-05
EIF4A2_amp	VEGF	639	-0.099	-5.349	9.1e-08	2.6e-05
WHSC1L1_amp	VEGF	438	-0.117	-5.343	9.4e-08	2.6e-05
CASP8_amp	Trail	73	0.295	5.318	1.1e-07	2.8e-05
SPRR3_amp	EGFR	432	-0.113	-5.317	1.1e-07	2.8e-05
ZC3H11A_amp	p53	332	0.076	5.323	1.1e-07	2.8e-05
CDK12_amp	NFkB	310	0.100	5.296	1.2e-07	3e-05
FGFR1_amp	VEGF	428	-0.117	-5.302	1.2e-07	2.9e-05
IRS2_amp	PI3K	173	-0.099	-5.289	1.3e-07	3e-05

MED24_amp	JAK-STAT	261	-0.093	-5.260	1.5e-07	3.4e-05
MYC_amp	PI3K	974	-0.045	-5.229	1.8e-07	4e-05
PTEN_del	p53	404	0.068	5.216	1.9e-07	4.2e-05
ABL2_amp	JAK-STAT	298	-0.087	-5.179	2.3e-07	4.9e-05
ARHGEF2_amp	EGFR	439	-0.110	-5.181	2.3e-07	4.9e-05
PIP5K1A_amp	EGFR	504	-0.103	-5.166	2.5e-07	5.1e-05
CDK12_amp	JAK-STAT	310	-0.083	-5.109	3.3e-07	6.7e-05
ING1_amp	PI3K	169	-0.096	-5.103	3.4e-07	6.8e-05
MED24_amp	Trail	261	0.152	5.097	3.5e-07	6.9e-05
PRRX1_amp	EGFR	339	-0.121	-5.093	3.6e-07	7e-05
SETDB1_amp	PI3K	513	-0.057	-5.071	4.1e-07	7.5e-05
ASH1L_amp	EGFR	451	-0.106	-5.072	4.1e-07	7.5e-05
ARHGEF2_amp	PI3K	439	-0.061	-5.058	4.4e-07	7.9e-05
RAD21_amp	p53	722	0.050	5.031	5e-07	9e-05
MECOM_amp	VEGF	783	-0.085	-4.991	6.2e-07	0.00011
PRRX1_amp	JAK-STAT	339	-0.079	-4.990	6.2e-07	0.00011
DHX9_amp	JAK-STAT	298	-0.084	-4.984	6.4e-07	0.00011
WHSC1L1_amp	JAK-STAT	438	-0.068	-4.970	6.9e-07	0.00011
FRG1B_amp	Trail	239	-0.153	-4.947	7.8e-07	0.00013
MUC20_amp	TGFb	647	-0.036	-4.937	8.1e-07	0.00013
ERBB2_amp	NFkB	378	0.085	4.934	8.2e-07	0.00013
NDRG1_amp	PI3K	744	-0.048	-4.937	8.2e-07	0.00013
MYC_amp	TNFa	974	0.045	4.932	8.4e-07	0.00013
TFDP1_amp	PI3K	178	-0.091	-4.918	8.9e-07	0.00013
FGFR1_amp	Trail	428	-0.115	-4.891	1e-06	0.00015
PTEN_del	TNFa	404	0.065	4.869	1.1e-06	0.00017
ASXL1_amp	MAPK	246	-0.038	-4.873	1.1e-06	0.00017
MED24_amp	NFkB	261	0.100	4.862	1.2e-06	0.00017
ARFGEF1_amp	NFkB	338	0.089	4.846	1.3e-06	0.00018
PABPC1_amp	JAK-STAT	664	-0.056	-4.846	1.3e-06	0.00018
PTEN_del	PI3K	404	-0.061	-4.843	1.3e-06	0.00018
ABL2_amp	EGFR	298	-0.122	-4.829	1.4e-06	0.00019
EIF4G1_amp	TGFb	665	-0.034	-4.810	1.5e-06	0.0002
CAPN7_del	Hypoxia	72	0.080	4.788	1.7e-06	0.00022
CHD1L_amp	PI3K	456	-0.057	-4.790	1.7e-06	0.00022
TXNIP_amp	PI3K	456	-0.057	-4.790	1.7e-06	0.00022
WHSC1L1_amp	Trail	438	-0.111	-4.795	1.7e-06	0.00022
FGFR1_amp	JAK-STAT	428	-0.066	-4.778	1.8e-06	0.00022
PIK3CA_amp	TGFb	735	-0.033	-4.779	1.8e-06	0.00022
CLASP2_del	Hypoxia	76	0.078	4.764	1.9e-06	0.00023
DLG1_amp	NFkB	654	0.064	4.765	1.9e-06	0.00023
FRG1B_amp	MAPK	239	-0.038	-4.758	2e-06	0.00024
VHL_del	Hypoxia	79	0.076	4.754	2e-06	0.00024
RAD21_amp	TNFa	722	0.049	4.736	2.2e-06	0.00026
MAP3K11_amp	Trail	117	0.207	4.717	2.4e-06	0.00028
PIK3C2B_amp	p53	355	0.065	4.716	2.5e-06	0.00028
RB1_del	TGFb	368	-0.043	-4.657	3.3e-06	0.00037
PIK3CA_amp	NFkB	735	0.059	4.642	3.5e-06	0.00039
FXR1_amp	TGFb	705	-0.032	-4.634	3.7e-06	0.0004
ARFGAP1_amp	JAK-STAT	279	-0.079	-4.621	3.9e-06	0.00042
TBL1XR1_amp	TGFb	710	-0.032	-4.620	3.9e-06	0.00042
CDC73_amp	JAK-STAT	280	-0.080	-4.603	4.3e-06	0.00045
NCF2_amp	JAK-STAT	307	-0.076	-4.596	4.4e-06	0.00046
NOTCH2_amp	PI3K	481	-0.053	-4.594	4.4e-06	0.00046
RB1_del	EGFR	368	-0.104	-4.577	4.8e-06	0.00049
DLG1_amp	TGFb	654	-0.033	-4.553	5.4e-06	0.00055
MED24_amp	Hypoxia	261	-0.041	-4.546	5.5e-06	0.00055
CAPN7_del	VEGF	72	0.240	4.541	5.7e-06	0.00055
CDK12_amp	Trail	310	0.125	4.539	5.7e-06	0.00055
RFC4_amp	TGFb	637	-0.033	-4.541	5.7e-06	0.00055
TBL1XR1_amp	NFkB	710	0.059	4.541	5.7e-06	0.00055
ELF3_amp	JAK-STAT	302	-0.075	-4.534	5.9e-06	0.00056
NCF2_amp	EGFR	307	-0.113	-4.534	5.9e-06	0.00056
NF1_del	PI3K	91	-0.117	-4.530	6e-06	0.00056
MYC_amp	JAK-STAT	974	-0.044	-4.522	6.3e-06	0.00058
MUC20_amp	NFkB	647	0.061	4.502	6.8e-06	0.00063
SPOP_amp	NFkB	168	0.115	4.497	7e-06	0.00064
ARFGAP1_amp	MAPK	279	-0.033	-4.489	7.3e-06	0.00066
RB1_del	MAPK	368	-0.029	-4.484	7.5e-06	0.00067
ACTG1_amp	PI3K	235	-0.072	-4.471	7.9e-06	0.0007
EIF4A2_amp	TGFb	639	-0.032	-4.455	8.5e-06	0.00075
DHX9_amp	EGFR	298	-0.113	-4.444	9e-06	0.00078
AHCTF1_amp	JAK-STAT	361	-0.068	-4.437	9.3e-06	0.0008
KEAP1_amp	TNFa	156	-0.092	-4.435	9.4e-06	0.0008
SPOP_amp	Trail	168	0.164	4.432	9.5e-06	0.0008

ARHGEF2_amp	JAK-STAT	439	-0.062	-4.425	9.8e-06	0.00083
ASH1L_amp	JAK-STAT	451	-0.061	-4.423	9.9e-06	0.00083
CTNNB1_del	Hypoxia	91	0.066	4.413	1e-05	0.00086
BPTF_amp	NFkB	274	0.089	4.409	1.1e-05	0.00086
FCRL4_amp	PI3K	368	-0.058	-4.406	1.1e-05	0.00086
MAP4K1_amp	Trail	259	-0.131	-4.408	1.1e-05	0.00086
MECOM_amp	NFkB	783	0.054	4.393	1.1e-05	0.0009
SPRR3_amp	PI3K	432	-0.054	-4.402	1.1e-05	0.00087
ASXL1_amp	Trail	246	-0.135	-4.410	1.1e-05	0.00086
ABL2_amp	MAPK	298	-0.032	-4.382	1.2e-05	0.00093
MNDA_amp	PI3K	370	-0.057	-4.373	1.2e-05	0.00096
MYC_amp	VEGF	974	-0.069	-4.386	1.2e-05	0.00092
NCF2_amp	MAPK	307	-0.031	-4.360	1.3e-05	0.001
ASPM_amp	JAK-STAT	279	-0.075	-4.362	1.3e-05	0.001
MLH1_del	Hypoxia	83	0.068	4.343	1.4e-05	0.0011
ATR_amp	VEGF	330	-0.109	-4.335	1.5e-05	0.0011
PPM1D_amp	NFkB	264	0.089	4.332	1.5e-05	0.0011
BMPR2_amp	Trail	68	0.248	4.312	1.6e-05	0.0012
PABPC1_amp	PI3K	664	-0.043	-4.300	1.7e-05	0.0013
PTEN_del	EGFR	404	-0.093	-4.305	1.7e-05	0.0012
SPOP_amp	VEGF	168	-0.149	-4.301	1.7e-05	0.0012
CCT5_amp	Hypoxia	406	0.032	4.295	1.8e-05	0.0013
DHX9_amp	MAPK	298	-0.031	-4.296	1.8e-05	0.0013
PRPF8_amp	Trail	40	-0.320	-4.265	2e-05	0.0014
SMARCA4_amp	TNFa	168	-0.086	-4.265	2e-05	0.0014
COL1A1_amp	NFkB	170	0.108	4.258	2.1e-05	0.0015
SRGAP3_del	Hypoxia	93	0.063	4.256	2.1e-05	0.0015
NDRG1_amp	TNFa	744	0.043	4.236	2.3e-05	0.0016
ARFGEF2_amp	JAK-STAT	226	-0.080	-4.228	2.4e-05	0.0016
VHL_amp	JAK-STAT	112	-0.112	-4.220	2.5e-05	0.0017
NEDD4L_amp	TNFa	41	-0.170	-4.208	2.6e-05	0.0017
PSME3_amp	TGFb	57	-0.097	-4.213	2.6e-05	0.0017
AXIN2_amp	NFkB	239	0.091	4.202	2.7e-05	0.0018

Table B8: Reactome vs. CNAs (pan-cancer)

CNA	Pathway	Size	Effect	Wald stat.	P-value	FDR
MSR1_del	Trail	392	-0.209	-9.748	2.8e-22	2.8e-18
CTTN_amp	TNFa	564	0.107	9.460	4.1e-21	2e-17
PIP5K1A_amp	TNFa	504	0.111	9.254	3e-20	9.9e-17
SETDB1_amp	TNFa	513	0.107	9.022	2.5e-19	6.1e-16
MYC_amp	p53	974	0.078	8.982	3.6e-19	7.2e-16
NOTCH2_amp	TNFa	481	0.106	8.727	3.3e-18	5.5e-15
CHD1L_amp	TNFa	456	0.106	8.518	2e-17	2.5e-14
TXNIP_amp	TNFa	456	0.106	8.518	2e-17	2.5e-14
SPRR3_amp	TNFa	432	0.108	8.411	5.1e-17	5.6e-14
PTEN_del	PI3K	404	-0.061	-8.324	1e-16	1e-13
ASH1L_amp	TNFa	451	0.104	8.245	2e-16	1.8e-13
FGFR1_amp	Trail	428	-0.168	-8.221	2.4e-16	2e-13
WHSC1L1_amp	Trail	438	-0.164	-8.082	7.5e-16	5.7e-13
ARHGEF2_amp	TNFa	439	0.103	8.072	8.4e-16	6e-13
NDRG1_amp	p53	744	0.078	8.007	1.4e-15	8.9e-13
PABPC1_amp	p53	664	0.082	8.013	1.4e-15	8.9e-13
CCND1_amp	TNFa	616	0.086	7.972	1.8e-15	1.1e-12
FCRL4_amp	TNFa	368	0.109	7.835	5.6e-15	3.1e-12
MNDA_amp	TNFa	370	0.107	7.733	1.2e-14	6.5e-12
RAD21_amp	p53	722	0.075	7.601	3.5e-14	1.7e-11
PABPC1_amp	Trail	664	-0.119	-6.971	3.5e-12	1.7e-09
MYC_amp	TGFb	974	0.049	6.942	4.3e-12	2e-09
CTTN_amp	Trail	564	0.125	6.882	6.4e-12	2.8e-09
CASP8_amp	Trail	73	0.330	6.787	1.2e-11	5.1e-09
SPOP_amp	p53	168	0.125	6.498	8.7e-11	3.5e-08
COL1A1_amp	p53	170	0.122	6.346	2.3e-10	9e-08
HLF_amp	p53	194	0.114	6.328	2.6e-10	9.7e-08
GATA3_amp	PI3K	155	-0.073	-6.317	2.8e-10	1e-07
SOX17_amp	Trail	319	-0.142	-5.978	2.4e-09	8.2e-07
DDX5_amp	p53	252	0.093	5.874	4.5e-09	1.5e-06
PPM1D_amp	p53	264	0.090	5.815	6.3e-09	2e-06
SMURF2_amp	p53	254	0.092	5.799	7e-09	2.2e-06
CDKN2A_del	TNFa	1060	0.050	5.769	8.4e-09	2.5e-06
BPTF_amp	p53	274	0.088	5.742	9.8e-09	2.9e-06
PTEN_del	EGFR	404	-0.046	-5.734	1e-08	2.9e-06
ARFGEF1_amp	p53	338	0.078	5.596	2.3e-08	6.4e-06

ARFGEF1_amp	Trail	338	-0.129	-5.568	2.7e-08	7.2e-06
PRRX1_amp	TNFa	339	0.081	5.563	2.8e-08	7.3e-06
AXIN2_amp	p53	239	0.089	5.488	4.2e-08	1.1e-05
MYC_amp	Hypoxia	974	0.042	5.482	4.4e-08	1.1e-05
MYC_amp	PI3K	974	-0.027	-5.454	5.2e-08	1.2e-05
RAD21_amp	PI3K	722	-0.031	-5.438	5.6e-08	1.3e-05
PRKAR1A_amp	p53	259	0.085	5.433	5.7e-08	1.3e-05
BMPR2_amp	Trail	68	0.270	5.357	8.7e-08	1.9e-05
PABPC1_amp	PI3K	664	-0.032	-5.359	8.7e-08	1.9e-05
FRG1B_amp	Trail	239	-0.145	-5.357	8.8e-08	1.9e-05
ASH1L_amp	PI3K	451	-0.038	-5.358	8.8e-08	1.9e-05
CLTC_amp	p53	252	0.085	5.327	1e-07	2.1e-05
NDRG1_amp	PI3K	744	-0.030	-5.296	1.2e-07	2.5e-05
CDKN2A_del	Trail	1060	0.074	5.286	1.3e-07	2.5e-05
KLF6_amp	PI3K	157	-0.061	-5.287	1.3e-07	2.5e-05
NDRG1_amp	TGFb	744	0.042	5.275	1.4e-07	2.6e-05
CCND1_amp	Trail	616	0.091	5.218	1.9e-07	3.5e-05
PIK3C2B_amp	TNFa	355	0.073	5.174	2.4e-07	4.4e-05
MED24_amp	p53	261	0.080	5.129	3e-07	5.4e-05
WHSC1L1_amp	JAK-STAT	438	-0.058	-5.115	3.2e-07	5.7e-05
ABL2_amp	VEGF	298	-0.061	-5.091	3.7e-07	6.4e-05
RAD21_amp	Hypoxia	722	0.044	5.082	3.9e-07	6.6e-05
NDRG1_amp	Hypoxia	744	0.044	5.064	4.2e-07	7.2e-05
ASH1L_amp	EGFR	451	-0.039	-5.045	4.7e-07	7.8e-05
VIM_amp	PI3K	91	-0.075	-5.020	5.3e-07	8.6e-05
CDKN2A_del	p53	1060	0.042	5.004	5.8e-07	9.2e-05
FMR1_amp	Hypoxia	124	0.098	4.985	6.3e-07	0.0001
MNDA_amp	PI3K	370	-0.039	-4.958	7.3e-07	0.00011
FGFR1_amp	JAK-STAT	428	-0.057	-4.949	7.6e-07	0.00012
FGFR1_amp	VEGF	428	-0.049	-4.941	7.9e-07	0.00012
KRAS_amp	TGFb	277	0.060	4.938	8.1e-07	0.00012
ARID4B_amp	VEGF	346	-0.055	-4.932	8.4e-07	0.00012
AFF4_amp	JAK-STAT	103	0.113	4.921	8.8e-07	0.00013
MECOM_amp	TGFb	783	0.037	4.921	8.8e-07	0.00013
PPM1D_amp	TGFb	264	0.062	4.913	9.2e-07	0.00013
SMAD4_del	PI3K	208	-0.049	-4.898	1e-06	0.00014
HLF_amp	TGFb	194	0.070	4.855	1.2e-06	0.00017
FCRL4_amp	PI3K	368	-0.038	-4.845	1.3e-06	0.00017
CHD1L_amp	PI3K	456	-0.034	-4.818	1.5e-06	0.00019
DHX9_amp	VEGF	298	-0.057	-4.810	1.5e-06	0.0002
TXNIP_amp	PI3K	456	-0.034	-4.818	1.5e-06	0.00019
PRRX1_amp	VEGF	339	-0.054	-4.806	1.6e-06	0.0002
WHSC1L1_amp	VEGF	438	-0.047	-4.793	1.7e-06	0.00021
ZC3H11A_amp	TNFa	332	0.070	4.792	1.7e-06	0.00021
PRKAR1A_amp	TGFb	259	0.060	4.775	1.8e-06	0.00023
MNDA_amp	EGFR	370	-0.040	-4.773	1.9e-06	0.00023
INPPL1_amp	TNFa	201	0.088	4.753	2e-06	0.00024
MNDA_amp	VEGF	370	-0.051	-4.723	2.4e-06	0.00028
MYC_amp	Trail	974	-0.069	-4.723	2.4e-06	0.00028
FCRL4_amp	EGFR	368	-0.040	-4.709	2.5e-06	0.00029
PSIP1_amp	JAK-STAT	92	0.113	4.712	2.5e-06	0.00029
SOX17_amp	p53	319	0.067	4.711	2.5e-06	0.00029
SPRR3_amp	EGFR	432	-0.037	-4.707	2.6e-06	0.00029
ASPM_amp	VEGF	279	-0.058	-4.704	2.6e-06	0.00029
BPTF_amp	TGFb	274	0.058	4.696	2.7e-06	0.00029
FCRL4_amp	VEGF	368	-0.051	-4.700	2.7e-06	0.00029
NCF2_amp	VEGF	307	-0.055	-4.654	3.3e-06	0.00035
NOTCH2_amp	PI3K	481	-0.032	-4.654	3.3e-06	0.00035
CHD1L_amp	VEGF	456	-0.045	-4.643	3.5e-06	0.00036
TXNIP_amp	VEGF	456	-0.045	-4.643	3.5e-06	0.00036
ARID4B_amp	EGFR	346	-0.040	-4.610	4.1e-06	0.00042
SPRR3_amp	VEGF	432	-0.046	-4.595	4.4e-06	0.00045
ELF3_amp	VEGF	302	-0.054	-4.586	4.6e-06	0.00046
ARHGEF2_amp	PI3K	439	-0.033	-4.588	4.6e-06	0.00046
RAD21_amp	Trail	722	-0.076	-4.574	4.9e-06	0.00048
ERBB2_amp	p53	378	0.060	4.564	5.1e-06	0.0005
SPRR3_amp	PI3K	432	-0.033	-4.554	5.4e-06	0.00052
ASH1L_amp	VEGF	451	-0.045	-4.551	5.4e-06	0.00052
HDAC3_amp	JAK-STAT	99	0.106	4.541	5.7e-06	0.00054
CHD1L_amp	EGFR	456	-0.034	-4.527	6.1e-06	0.00056
PPP2R5A_amp	TNFa	259	0.074	4.527	6.1e-06	0.00056
TXNIP_amp	EGFR	456	-0.034	-4.527	6.1e-06	0.00056
DIS3_del	NFkB	123	-0.100	-4.502	6.8e-06	0.00062
SETDB1_amp	VEGF	513	-0.041	-4.461	8.3e-06	0.00075
AFF4_amp	Trail	103	0.184	4.456	8.5e-06	0.00076

F8_amp	NFkB	176	0.083	4.453	8.6e-06	0.00076
TBL1XR1_amp	p53	710	0.044	4.452	8.7e-06	0.00076
PABPC1_amp	Hypoxia	664	0.040	4.427	9.7e-06	0.00085
PIP5K1A_amp	VEGF	504	-0.041	-4.420	1e-05	0.00086
ARHGEF2_amp	EGFR	439	-0.034	-4.417	1e-05	0.00087
ARHGEF6_amp	Hypoxia	89	0.103	4.420	1e-05	0.00086
CDKN2A_del	PI3K	1060	-0.021	-4.391	1.1e-05	0.00095
PIP5K1A_amp	EGFR	504	-0.032	-4.400	1.1e-05	0.00092
SETDB1_amp	PI3K	513	-0.029	-4.407	1.1e-05	0.0009
CLTC_amp	TGFb	252	0.056	4.376	1.2e-05	0.00099
ARAP3_amp	JAK-STAT	100	0.102	4.380	1.2e-05	0.00098
NDRG1_amp	Trail	744	-0.072	-4.379	1.2e-05	0.00098
PIK3C2B_amp	VEGF	355	-0.048	-4.384	1.2e-05	0.00098
CDK12_amp	p53	310	0.063	4.356	1.3e-05	0.001
COL1A1_amp	NFkB	170	-0.083	-4.368	1.3e-05	0.00099
FIP1L1_amp	Trail	139	-0.154	-4.360	1.3e-05	0.001
KDR_amp	Trail	116	-0.169	-4.369	1.3e-05	0.00099
NOTCH2_amp	VEGF	481	-0.041	-4.362	1.3e-05	0.001
SETDB1_amp	EGFR	513	-0.032	-4.371	1.3e-05	0.00099
SMURF2_amp	TGFb	254	0.056	4.368	1.3e-05	0.00099
TCF4_amp	NFkB	32	-0.188	-4.361	1.3e-05	0.001
KDM6A_del	EGFR	135	-0.059	-4.352	1.4e-05	0.001
STK11_del	PI3K	91	-0.066	-4.344	1.4e-05	0.0011
CDC73_amp	VEGF	280	-0.053	-4.329	1.5e-05	0.0011
HCFC1_amp	NFkB	195	0.077	4.311	1.6e-05	0.0012
WNK1_amp	TGFb	267	0.053	4.312	1.6e-05	0.0012
EIF4G1_amp	NFkB	665	-0.043	-4.311	1.7e-05	0.0012
KDM6A_del	PI3K	135	-0.054	-4.308	1.7e-05	0.0012
ARID4B_amp	PI3K	346	-0.035	-4.311	1.7e-05	0.0012
PIP5K1A_amp	PI3K	504	-0.029	-4.285	1.9e-05	0.0013
ARFGEF1_amp	Hypoxia	338	0.052	4.263	2e-05	0.0014
NOTCH2_amp	EGFR	481	-0.032	-4.270	2e-05	0.0014
ARHGEF2_amp	VEGF	439	-0.042	-4.265	2e-05	0.0014
ZC3H11A_amp	VEGF	332	-0.048	-4.263	2e-05	0.0014
NFE2L2_amp	p53	133	0.092	4.259	2.1e-05	0.0014
DDX5_amp	NFkB	252	-0.067	-4.251	2.2e-05	0.0015
FXR1_amp	NFkB	705	-0.041	-4.250	2.2e-05	0.0015
MCM8_amp	Trail	85	-0.190	-4.245	2.2e-05	0.0015
DDX5_amp	TGFb	252	0.054	4.240	2.3e-05	0.0015

Table B9: SPIA vs. CNAs (pan-cancer)

CNA	Pathway	Size	Effect	Wald stat.	P-value	FDR
MSR1_del	Trail	341	-21.244	-13.496	1.5e-40	1.1e-36
HDAC3_amp	NFkB	87	43.874	12.053	5e-33	1.2e-29
ARAP3_amp	NFkB	87	43.872	12.052	5e-33	1.2e-29
PIK3CA_amp	MAPK	621	-15.399	-11.950	2e-32	3.6e-29
FXR1_amp	MAPK	595	-15.554	-11.835	7.7e-32	9.3e-29
CLASP2_del	NFkB	69	45.058	11.843	7.8e-32	9.3e-29
TBL1XR1_amp	MAPK	595	-15.399	-11.716	3e-31	3.1e-28
EIF4G1_amp	MAPK	561	-15.678	-11.639	7.3e-31	6.6e-28
CLASP2_del	VEGF	69	19.588	11.570	1.8e-30	1.4e-27
MLH1_del	VEGF	77	18.555	11.558	2e-30	1.5e-27
CSNK1G3_amp	NFkB	78	44.343	11.501	2.9e-30	1.7e-27
ARHGAP26_amp	NFkB	94	40.340	11.500	3e-30	1.7e-27
AFF4_amp	NFkB	90	41.309	11.494	3.1e-30	1.7e-27
MECOM_amp	MAPK	643	-14.604	-11.503	3.4e-30	1.8e-27
MYD88_del	NFkB	68	43.618	11.366	1.7e-29	8.4e-27
ACSL6_amp	NFkB	86	41.353	11.252	4.7e-29	2.1e-26
NPM1_amp	NFkB	124	34.537	11.248	5.1e-29	2.1e-26
G3BP1_amp	NFkB	91	39.829	11.167	1.2e-28	4.9e-26
PPM1D_amp	PI3K	222	-3.518	-11.164	1.4e-28	5.2e-26
FAT2_amp	NFkB	92	39.180	11.043	4.7e-28	1.7e-25
CLTC_amp	PI3K	212	-3.552	-10.998	8.3e-28	2.9e-25
CAPN7_del	NFkB	69	41.875	10.973	1.2e-27	4e-25
CTNNB1_del	VEGF	78	17.413	10.873	3.7e-27	1.2e-24
MLH1_del	NFkB	77	39.206	10.850	4.7e-27	1.4e-24
MYD88_del	VEGF	68	18.553	10.839	5.3e-27	1.5e-24
CTNNB1_del	NFkB	78	38.705	10.788	9.2e-27	2.5e-24
PABPC1_amp	Trail	523	-14.336	-10.713	2.2e-26	5.9e-24
CAPN7_del	VEGF	69	18.071	10.678	2.8e-26	7.3e-24
ITGA9_del	VEGF	77	16.964	10.532	1.3e-25	3.3e-23
RFC4_amp	MAPK	537	-14.271	-10.380	5.9e-25	1.4e-22

EIF4A2_amp	MAPK	538	-14.231	-10.360	7.2e-25	1.7e-22
HDAC3_amp	VEGF	87	15.636	10.314	1e-24	2.2e-22
ARAP3_amp	VEGF	87	15.635	10.313	1e-24	2.2e-22
ITGA9_del	NFkB	77	37.400	10.329	1.1e-24	2.2e-22
SETDB1_amp	MAPK	432	-15.925	-10.332	1.1e-24	2.2e-22
NSD1_amp	NFkB	150	28.730	10.242	2.2e-24	4.3e-22
ASH1L_amp	MAPK	364	-16.963	-10.185	4.7e-24	9.2e-22
VHL_del	NFkB	76	37.062	10.170	5.1e-24	9.7e-22
PIP5K1A_amp	MAPK	420	-15.741	-10.077	1.4e-23	2.5e-21
CLASP2_del	Trail	69	32.334	10.017	2.4e-23	4.3e-21
G3BP1_amp	VEGF	91	14.814	9.974	3.2e-23	5.6e-21
RHOA_del	VEGF	103	13.777	9.863	1.1e-22	1.9e-20
ARHGAP26_amp	VEGF	94	14.398	9.837	1.2e-22	2e-20
FGFR1_amp	Trail	389	-15.125	-9.825	1.4e-22	2.3e-20
AFF4_amp	VEGF	90	14.671	9.810	1.6e-22	2.5e-20
ACSL6_amp	VEGF	86	14.992	9.801	1.7e-22	2.7e-20
ARHGEF2_amp	MAPK	357	-16.388	-9.739	3.7e-22	5.7e-20
WHSC1L1_amp	Trail	399	-14.672	-9.632	9.1e-22	1.4e-19
SETD2_del	VEGF	97	13.879	9.636	9.8e-22	1.4e-19
TGFBR2_del	Trail	98	26.311	9.628	1e-21	1.5e-19
TGFBR2_del	NFkB	98	30.982	9.625	1.1e-21	1.5e-19
SPRR3_amp	MAPK	360	-16.076	-9.619	1.2e-21	1.6e-19
FAT2_amp	VEGF	92	14.173	9.590	1.3e-21	1.8e-19
MNDA_amp	MAPK	315	-17.178	-9.599	1.4e-21	1.9e-19
MYD88_del	Trail	68	31.170	9.585	1.6e-21	2.1e-19
TGFBR2_del	VEGF	98	13.692	9.562	1.9e-21	2.5e-19
DLG1_amp	MAPK	554	-12.964	-9.553	2e-21	2.6e-19
MUC20_amp	MAPK	547	-12.972	-9.497	3.4e-21	4.2e-19
ITGA9_del	Trail	77	29.126	9.501	3.5e-21	4.2e-19
VHL_del	VEGF	76	15.316	9.484	4e-21	4.8e-19
RAD21_amp	Trail	540	-12.483	-9.488	4.2e-21	5e-19
FCRL4_amp	MAPK	309	-17.013	-9.457	5.4e-21	6.3e-19
CDKN2A_del	NFkB	793	-12.543	-9.423	6.8e-21	7.8e-19
MLH1_del	Trail	77	28.769	9.407	8.3e-21	9.4e-19
APC_amp	NFkB	63	40.288	9.362	1.1e-20	1.2e-18
PBRM1_del	VEGF	110	12.642	9.303	2.2e-20	2.4e-18
ASH1L_amp	Trail	364	-15.068	-9.289	2.6e-20	2.7e-18
RBM5_del	VEGF	104	12.938	9.282	2.7e-20	2.9e-18
SETDB1_amp	Trail	432	-13.653	-9.093	1.5e-19	1.6e-17
SRGAP3_del	NFkB	85	31.354	9.058	2e-19	2.1e-17
CTNNB1_del	Trail	78	27.552	9.038	2.4e-19	2.5e-17
CSNK1G3_amp	VEGF	78	14.474	9.008	2.9e-19	2.9e-17
NPM1_amp	VEGF	124	11.525	8.998	3.1e-19	3.1e-17
NSD1_amp	VEGF	150	10.507	8.999	3.1e-19	3.1e-17
CAPN7_del	Trail	69	28.872	8.887	9.2e-19	8.8e-17
RHOA_del	NFkB	103	28.044	8.888	9.4e-19	8.9e-17
MYC_amp	Trail	742	-10.290	-8.846	1.4e-18	1.3e-16
PRRX1_amp	MAPK	284	-16.546	-8.783	2.4e-18	2.2e-16
SRGAP3_del	VEGF	85	13.436	8.779	2.4e-18	2.2e-16
NOTCH2_amp	MAPK	411	-13.870	-8.749	3.1e-18	2.8e-16
SETD2_del	NFkB	97	28.296	8.724	3.9e-18	3.5e-16
SMURF2_amp	PI3K	211	-2.810	-8.685	5.2e-18	4.5e-16
WHSC1L1_amp	NFkB	399	-15.918	-8.674	5.6e-18	4.9e-16
MECOM_amp	JAK-STAT	643	-1.597	-8.674	5.8e-18	4.9e-16
CHD1L_amp	MAPK	390	-14.014	-8.644	7.6e-18	6.3e-16
TXNIP_amp	MAPK	390	-14.014	-8.644	7.6e-18	6.3e-16
PIP5K1A_amp	Trail	420	-13.122	-8.617	9.9e-18	8.2e-16
CLASP2_del	MAPK	69	28.957	8.600	1.1e-17	9.2e-16
RBM5_del	NFkB	104	27.021	8.601	1.1e-17	9.2e-16
FGFR1_amp	NFkB	389	-15.903	-8.570	1.4e-17	1.1e-15
BCL11A_amp	MAPK	82	-28.297	-8.550	1.6e-17	1.3e-15
BPTF_amp	PI3K	223	-2.689	-8.544	1.7e-17	1.4e-15
ARHGEF2_amp	Trail	357	-13.998	-8.550	1.8e-17	1.4e-15
DDX5_amp	PI3K	209	-2.776	-8.530	1.9e-17	1.5e-15
FGFR1_amp	PI3K	389	-2.074	-8.511	2.3e-17	1.7e-15
RASA2_amp	MAPK	276	-15.536	-8.456	3.7e-17	2.8e-15
BAP1_del	VEGF	109	11.572	8.462	3.7e-17	2.8e-15
FCRL4_amp	Trail	309	-14.723	-8.405	6e-17	4.4e-15
SPRR3_amp	Trail	360	-13.656	-8.397	6.4e-17	4.6e-15
CDKN2A_del	PI3K	793	1.475	8.389	6.5e-17	4.6e-15
NDRG1_amp	Trail	542	-11.103	-8.398	6.5e-17	4.6e-15
APC_amp	VEGF	63	14.967	8.357	8e-17	5.7e-15
FGFR1_amp	MAPK	389	-13.285	-8.340	9.6e-17	6.7e-15
WHSC1L1_amp	PI3K	399	-1.999	-8.300	1.3e-16	9.2e-15
FXR1_amp	JAK-STAT	595	-1.581	-8.299	1.4e-16	9.4e-15

NPM1_amp	PI3K	124	-3.466	-8.274	1.6e-16	1.1e-14
ARAP3_amp	PI3K	87	-4.122	-8.263	1.8e-16	1.2e-14
HDAC3_amp	PI3K	87	-4.120	-8.257	1.9e-16	1.2e-14
ARHGAP26_amp	PI3K	94	-3.968	-8.254	1.9e-16	1.3e-14
ACSL6_amp	PI3K	86	-4.142	-8.246	2e-16	1.3e-14
SETD2_del	Trail	97	22.570	8.235	2.4e-16	1.6e-14
CSNK1G3_amp	PI3K	78	-4.331	-8.220	2.5e-16	1.6e-14
CAPN7_del	MAPK	69	27.702	8.219	2.7e-16	1.7e-14
G3BP1_amp	PI3K	91	-4.005	-8.209	2.8e-16	1.7e-14
PIK3CA_amp	JAK-STAT	621	-1.537	-8.211	2.8e-16	1.8e-14
PRRX1_amp	Trail	284	-14.998	-8.217	2.9e-16	1.8e-14
WHSC1L1_amp	MAPK	399	-12.921	-8.203	3e-16	1.8e-14
PBRM1_del	NFkB	110	25.065	8.208	3.1e-16	1.9e-14
AFF4_amp	MAPK	90	26.357	8.187	3.3e-16	2e-14
EIF4G1_amp	NFkB	561	-12.787	-8.188	3.4e-16	2.1e-14
TFDP2_amp	MAPK	287	-14.760	-8.181	3.6e-16	2.2e-14
TBL1XR1_amp	JAK-STAT	595	-1.553	-8.154	4.5e-16	2.7e-14
MLH1_del	MAPK	77	25.958	8.151	4.8e-16	2.8e-14
FAT2_amp	PI3K	92	-3.948	-8.133	5.1e-16	3e-14
FXR1_amp	NFkB	595	-12.415	-8.141	5.1e-16	2.9e-14
COL1A1_amp	PI3K	150	-3.119	-8.135	5.2e-16	3e-14
PRKAR1A_amp	PI3K	207	-2.660	-8.127	5.6e-16	3.2e-14
MNDA_amp	Trail	315	-14.110	-8.096	7.6e-16	4.3e-14
NSD1_amp	PI3K	150	-3.092	-8.086	7.6e-16	4.3e-14
XPO1_amp	MAPK	82	-26.742	-8.075	8.4e-16	4.7e-14
HDAC3_amp	MAPK	87	26.418	8.065	9e-16	4.9e-14
ARAP3_amp	MAPK	87	26.411	8.063	9.1e-16	5e-14
FOXA1_amp	MAPK	142	-20.268	-8.050	1e-15	5.5e-14
ABL2_amp	Trail	249	-15.636	-8.051	1.1e-15	5.9e-14
PIK3CB_amp	MAPK	255	-15.319	-8.019	1.3e-15	7.2e-14
MYD88_del	MAPK	68	27.188	8.014	1.4e-15	7.6e-14
HLF_amp	PI3K	165	-2.934	-8.008	1.5e-15	7.6e-14
XRN1_amp	MAPK	289	-14.379	-7.977	1.9e-15	9.9e-14
CHD1L_amp	Trail	390	-12.510	-7.949	2.4e-15	1.2e-13
TXNIP_amp	Trail	390	-12.510	-7.949	2.4e-15	1.2e-13
ATR_amp	MAPK	289	-14.309	-7.945	2.4e-15	1.2e-13
BAP1_del	NFkB	109	24.346	7.937	2.7e-15	1.4e-13
ABL2_amp	MAPK	249	-15.825	-7.931	2.8e-15	1.4e-13
SPOP_amp	PI3K	145	-3.080	-7.900	3.4e-15	1.7e-13
TGFBR2_del	MAPK	98	22.330	7.887	4e-15	2e-13
BPTF_amp	MAPK	223	-15.958	-7.873	4.3e-15	2.1e-13
RB1_del	MAPK	272	-14.594	-7.871	4.4e-15	2.2e-13
RHOA_del	Trail	103	20.930	7.868	4.6e-15	2.3e-13
EIF4G1_amp	JAK-STAT	561	-1.519	-7.785	8.6e-15	4.2e-13
WNT5A_del	VEGF	92	11.600	7.788	8.7e-15	4.2e-13

Table B10: Pathifier vs. CNAs (pan-cancer)

CNA	Pathway	Size	Effect	Wald stat.	P-value	FDR
PIK3CA_amp	VEGF	621	0.102	12.429	6.8e-35	4.5e-31
EIF4G1_amp	NFkB	561	0.124	12.406	9e-35	4.5e-31
EIF4G1_amp	VEGF	561	0.105	12.273	4.5e-34	1e-30
FXR1_amp	VEGF	595	0.102	12.265	4.9e-34	1e-30
PIK3CA_amp	NFkB	621	0.117	12.259	5.2e-34	1e-30
FXR1_amp	NFkB	595	0.119	12.228	7.6e-34	1.3e-30
TBL1XR1_amp	NFkB	595	0.118	12.065	5.3e-33	7.5e-30
TBL1XR1_amp	VEGF	595	0.100	12.016	9.4e-33	1.2e-29
MUC20_amp	VEGF	547	0.104	11.966	1.7e-32	1.8e-29
CTTN_amp	Trail	529	0.124	11.801	9.7e-32	9.6e-29
DLG1_amp	VEGF	554	0.101	11.732	2.5e-31	2.3e-28
MECOM_amp	VEGF	643	0.094	11.640	7.3e-31	6e-28
RFC4_amp	VEGF	537	0.101	11.562	1.7e-30	1.3e-27
EIF4A2_amp	VEGF	538	0.100	11.497	3.6e-30	2.6e-27
MECOM_amp	NFkB	643	0.109	11.434	7.4e-30	4.9e-27
DLG1_amp	NFkB	554	0.114	11.392	1.2e-29	7.2e-27
CTTN_amp	TNFa	529	0.102	11.334	1.9e-29	1.1e-26
RFC4_amp	NFkB	537	0.115	11.306	3.1e-29	1.7e-26
EIF4A2_amp	NFkB	538	0.115	11.277	4.2e-29	2.2e-26
MUC20_amp	NFkB	547	0.114	11.263	4.9e-29	2.4e-26
TBL1XR1_amp	EGFR	595	0.085	11.234	6.8e-29	3.2e-26
PIK3CA_amp	EGFR	621	0.083	11.162	1.5e-28	6.7e-26
CCND1_amp	TNFa	581	0.095	11.020	6.1e-28	2.6e-25
DLG1_amp	EGFR	554	0.086	10.951	1.5e-27	6e-25
FXR1_amp	EGFR	595	0.082	10.772	9.9e-27	3.9e-24

MUC20_amp	EGFR	547	0.085	10.754	1.2e-26	4.6e-24
CCND1_amp	Trail	581	0.106	10.444	2.7e-25	1e-22
EIF4G1_amp	EGFR	561	0.080	10.357	7.4e-25	2.6e-22
FXR1_amp	MAPK	595	0.079	9.923	5.7e-23	2e-20
PIK3CA_amp	MAPK	621	0.078	9.909	6.6e-23	2.2e-20
RFC4_amp	EGFR	537	0.078	9.872	9.5e-23	3e-20
EIF4A2_amp	EGFR	538	0.078	9.852	1.1e-22	3.6e-20
TBL1XR1_amp	MAPK	595	0.077	9.607	1.2e-21	3.6e-19
EIF4G1_amp	MAPK	561	0.078	9.525	2.6e-21	7.7e-19
MECOM_amp	EGFR	643	0.070	9.498	3.4e-21	9.6e-19
DLG1_amp	MAPK	554	0.077	9.345	1.4e-20	3.9e-18
TRIO_amp	NFkB	365	0.114	9.275	2.7e-20	7.4e-18
MUC20_amp	MAPK	547	0.076	9.146	8.8e-20	2.3e-17
RFC4_amp	MAPK	537	0.074	8.832	1.5e-18	3.7e-16
EIF4A2_amp	MAPK	538	0.074	8.809	1.8e-18	4.4e-16
XRN1_amp	EGFR	289	0.091	8.669	6e-18	1.4e-15
ARHGAP26_amp	JAK-STAT	94	0.180	8.646	6.9e-18	1.6e-15
ZNF292_del	NFkB	118	-0.179	-8.606	1e-17	2.4e-15
CCT5_amp	NFkB	342	0.108	8.550	1.7e-17	3.8e-15
TFDP2_amp	EGFR	287	0.090	8.524	2.1e-17	4.6e-15
CLASP2_del	p53	69	0.225	8.490	2.9e-17	6.2e-15
TRIO_amp	Trail	365	-0.105	-8.458	3.7e-17	7.8e-15
ATR_amp	VEGF	289	0.098	8.450	3.9e-17	8.1e-15
ATR_amp	EGFR	289	0.089	8.426	4.8e-17	9.6e-15
AFF4_amp	JAK-STAT	90	0.178	8.389	6.2e-17	1.2e-14
XRN1_amp	VEGF	289	0.097	8.322	1.1e-16	2.2e-14
BRWD1_del	NFkB	83	-0.204	-8.317	1.2e-16	2.2e-14
HDAC3_amp	JAK-STAT	87	0.179	8.302	1.3e-16	2.4e-14
ARAP3_amp	JAK-STAT	87	0.179	8.301	1.3e-16	2.4e-14
RASA2_amp	VEGF	276	0.098	8.245	2.1e-16	3.9e-14
TFDP2_amp	VEGF	287	0.096	8.212	2.8e-16	5e-14
ACSL6_amp	JAK-STAT	86	0.178	8.189	3.3e-16	5.7e-14
CCT5_amp	Trail	342	-0.105	-8.182	3.7e-16	6.3e-14
MYD88_del	p53	68	0.219	8.180	3.8e-16	6.3e-14
CSNK1G3_amp	p53	78	0.202	8.129	5.3e-16	8.8e-14
MECOM_amp	MAPK	643	0.063	8.125	5.7e-16	9.3e-14
MLH1_del	p53	77	0.202	8.036	1.2e-15	1.9e-13
TFDP2_amp	NFkB	287	0.109	8.021	1.3e-15	2.1e-13
XRN1_amp	NFkB	289	0.108	7.986	1.8e-15	2.7e-13
CSNK1G3_amp	JAK-STAT	78	0.182	7.962	2e-15	3.1e-13
ATR_amp	NFkB	289	0.107	7.951	2.3e-15	3.5e-13
RASA2_amp	EGFR	276	0.085	7.940	2.5e-15	3.8e-13
PIK3CB_amp	VEGF	255	0.097	7.893	3.7e-15	5.4e-13
HDAC3_amp	p53	87	0.184	7.787	8.2e-15	1.2e-12
ARAP3_amp	p53	87	0.184	7.787	8.2e-15	1.2e-12
ARHGAP26_amp	p53	94	0.177	7.785	8.3e-15	1.2e-12
ACSL6_amp	p53	86	0.184	7.766	9.6e-15	1.3e-12
RASA2_amp	NFkB	276	0.107	7.770	9.7e-15	1.3e-12
CDKN2A_del	VEGF	793	0.058	7.729	1.3e-14	1.8e-12
SYNCRIP_del	NFkB	104	-0.171	-7.725	1.4e-14	1.8e-12
EIF4G1_amp	TGFb	561	0.064	7.711	1.5e-14	2e-12
CTNNB1_del	p53	78	0.192	7.667	2.2e-14	2.8e-12
CAPN7_del	p53	69	0.201	7.616	3.2e-14	4.1e-12
AFF4_amp	Trail	90	0.188	7.592	3.7e-14	4.6e-12
CTTN_amp	VEGF	529	0.066	7.566	4.5e-14	5.6e-12
ITGA9_del	p53	77	0.189	7.503	7.7e-14	9.4e-12
CTTN_amp	EGFR	529	0.059	7.431	1.2e-13	1.5e-11
PIK3CA_amp	TGFb	621	0.059	7.339	2.5e-13	3e-11
TGFBR2_del	p53	98	0.164	7.344	2.5e-13	3e-11
FXR1_amp	TGFb	595	0.060	7.331	2.7e-13	3.2e-11
AFF4_amp	p53	90	0.170	7.309	3.1e-13	3.5e-11
CCND1_amp	VEGF	581	0.061	7.269	4.1e-13	4.7e-11
FAT2_amp	JAK-STAT	92	0.153	7.250	4.7e-13	5.3e-11
PIK3CB_amp	NFkB	255	0.104	7.246	5e-13	5.6e-11
G3BP1_amp	JAK-STAT	91	0.153	7.220	5.9e-13	6.5e-11
PIK3CB_amp	EGFR	255	0.080	7.131	1.2e-12	1.3e-10
RHOA_del	p53	103	0.157	7.130	1.2e-12	1.3e-10
LCP1_del	NFkB	167	-0.125	-7.079	1.7e-12	1.8e-10
XPO1_amp	NFkB	82	0.174	7.071	1.8e-12	1.8e-10
PBRM1_del	p53	110	0.151	7.067	1.9e-12	2e-10
RFC4_amp	TGFb	537	0.060	7.029	2.4e-12	2.5e-10
EIF4A2_amp	TGFb	538	0.060	7.004	2.9e-12	2.9e-10
TBL1XR1_amp	TGFb	595	0.056	6.909	5.6e-12	5.6e-10
STAG1_amp	VEGF	227	0.090	6.897	6.1e-12	6.1e-10
FAT2_amp	p53	92	0.158	6.877	6.8e-12	6.7e-10

G3BP1_amp	p53	91	0.159	6.873	7e-12	6.9e-10
MUC20_amp	TGFb	547	0.058	6.840	9e-12	8.7e-10
IRF6_amp	PI3K	245	0.098	6.819	1.1e-11	1e-09
PIK3C2B_amp	PI3K	280	0.092	6.816	1.1e-11	1e-09
CCND1_amp	EGFR	581	0.052	6.798	1.2e-11	1.1e-09
ACSL6_amp	Trail	86	0.172	6.778	1.3e-11	1.2e-09
CDKN2A_del	EGFR	793	0.047	6.791	1.3e-11	1.2e-09
BAP1_del	p53	109	0.145	6.756	1.6e-11	1.5e-09
CDKN2A_del	MAPK	793	0.049	6.697	2.4e-11	2.2e-09
STAG1_amp	EGFR	227	0.079	6.654	3.2e-11	2.9e-09
BCL11A_amp	NFkB	82	0.163	6.622	3.9e-11	3.5e-09
MLH1_del	JAK-STAT	77	0.158	6.627	3.9e-11	3.5e-09
FGFR1_amp	Trail	389	-0.081	-6.616	4.1e-11	3.6e-09
TRIO_amp	VEGF	365	0.070	6.609	4.3e-11	3.8e-09
NCK1_amp	VEGF	227	0.086	6.602	4.5e-11	3.9e-09
ACSL6_amp	PI3K	86	0.155	6.597	4.6e-11	3.9e-09
HDAC3_amp	Trail	87	0.165	6.543	6.6e-11	5.6e-09
ARAP3_amp	Trail	87	0.165	6.538	6.8e-11	5.7e-09
ARHGAP26_amp	Trail	94	0.159	6.531	7.1e-11	5.9e-09
ZC3H11A_amp	PI3K	265	0.090	6.492	9.5e-11	7.9e-09
G3BP1_amp	PI3K	91	0.148	6.473	1e-10	8.5e-09
WHSC1L1_amp	Trail	399	-0.078	-6.475	1e-10	8.5e-09
LCP1_del	p53	167	-0.116	-6.458	1.2e-10	9.5e-09
RBM5_del	p53	104	0.142	6.458	1.2e-10	9.5e-09
HDAC3_amp	PI3K	87	0.150	6.434	1.3e-10	1.1e-08
DLG1_amp	TGFb	554	0.054	6.429	1.4e-10	1.1e-08
ARAP3_amp	PI3K	87	0.150	6.433	1.4e-10	1.1e-08
VHL_del	p53	76	0.162	6.430	1.4e-10	1.1e-08
CLASP2_del	JAK-STAT	69	0.162	6.418	1.5e-10	1.2e-08
NCK1_amp	EGFR	227	0.076	6.406	1.6e-10	1.3e-08
CSNK1G3_amp	Trail	78	0.169	6.356	2.2e-10	1.7e-08
SETD2_del	p53	97	0.145	6.361	2.2e-10	1.7e-08
ARHGAP26_amp	PI3K	94	0.142	6.310	3e-10	2.3e-08
APC_amp	JAK-STAT	63	0.160	6.306	3.1e-10	2.3e-08
CSNK1G3_amp	PI3K	78	0.155	6.293	3.3e-10	2.4e-08
ACVR2A_del	NFkB	55	-0.189	-6.298	3.3e-10	2.4e-08
ASH1L_amp	PI3K	364	0.075	6.286	3.6e-10	2.6e-08
MECOM_amp	Trail	643	-0.062	-6.267	4e-10	2.9e-08
XRN1_amp	MAPK	289	0.069	6.240	4.8e-10	3.4e-08
CTNNB1_del	NFkB	78	0.161	6.217	5.6e-10	3.9e-08
KALRN_amp	NFkB	143	0.116	6.185	6.8e-10	4.8e-08
ELF1_del	NFkB	155	-0.113	-6.183	6.9e-10	4.8e-08
CTNNB1_del	JAK-STAT	78	0.147	6.179	7.1e-10	4.9e-08
CAPN7_del	NFkB	69	0.168	6.176	7.2e-10	4.9e-08
ATR_amp	MAPK	289	0.069	6.173	7.3e-10	5e-08
AFF4_amp	PI3K	90	0.142	6.155	8e-10	5.4e-08
STAG1_amp	NFkB	227	0.093	6.152	8.3e-10	5.6e-08
FAT2_amp	PI3K	92	0.139	6.132	9.3e-10	6.2e-08
EIF4G1_amp	Trail	561	-0.064	-6.125	9.8e-10	6.5e-08
MYD88_del	JAK-STAT	68	0.156	6.122	1e-09	6.7e-08

Table B11: PARADIGM vs. CNAs (pan-cancer)

CNA	Pathway	Size	Effect	Wald stat.	P-value	FDR
PIK3CA_amp	PI3K	732	1.667	24.349	3.5e-125	3.1e-121
TBL1XR1_amp	PI3K	708	1.642	23.585	7.1e-118	3.2e-114
FXR1_amp	PI3K	703	1.627	23.271	6e-115	1.8e-111
MECOM_amp	PI3K	781	1.493	22.263	1e-105	2.3e-102
EIF4G1_amp	PI3K	664	1.575	21.886	2.3e-102	4.2e-99
RFC4_amp	PI3K	636	1.527	20.747	1.6e-92	2.4e-89
EIF4A2_amp	PI3K	638	1.525	20.741	1.8e-92	2.4e-89
DLG1_amp	PI3K	652	1.430	19.576	7e-83	7.9e-80
MUC20_amp	PI3K	645	1.427	19.440	8.6e-82	8.7e-79
XRN1_amp	PI3K	329	1.429	14.141	1e-44	8.6e-42
ATR_amp	PI3K	329	1.428	14.139	1.1e-44	8.6e-42
TFDP2_amp	PI3K	327	1.415	13.957	1.3e-43	9.6e-41
RASA2_amp	PI3K	313	1.402	13.541	3.4e-41	2.4e-38
PIK3CB_amp	PI3K	292	1.338	12.470	2.8e-35	1.8e-32
NCK1_amp	PI3K	257	1.341	11.739	1.7e-31	1e-28
STAG1_amp	PI3K	257	1.341	11.734	1.8e-31	1e-28
KALRN_amp	PI3K	173	1.160	8.331	9.7e-17	5.2e-14
PLXNA1_amp	PI3K	153	1.195	8.098	6.7e-16	3.3e-13
ARFGAP1_del	VEGF	5	0.195	7.565	4.3e-14	2e-11
COL1A1_del	EGFR	5	0.267	6.007	2e-09	8.9e-07

CTCF_del	EGFR	59	0.074	5.669	1.5e-08	6.5e-06
SMARCB1_del	Trail	15	0.067	5.331	1e-07	4.2e-05
CDKN2A_del	JAK-STAT	1056	-0.253	-5.320	1.1e-07	4.2e-05
STK4_amp	MAPK	212	0.364	5.244	1.6e-07	6.1e-05
TNPO2_amp	PI3K	123	0.792	4.817	1.5e-06	0.00054
CEBPA_amp	PI3K	254	0.559	4.807	1.6e-06	0.00054
ASH1L_amp	JAK-STAT	449	-0.331	-4.804	1.6e-06	0.00054
RB1_del	TGFb	367	-0.382	-4.758	2e-06	0.00065
CDKN2A_del	TNFa	1056	-0.014	-4.688	2.8e-06	0.00088
ABL2_amp	JAK-STAT	296	-0.389	-4.647	3.5e-06	0.001
EPC1_del	VEGF	16	0.069	4.587	4.6e-06	0.0013
FOXA2_del	Trail	16	0.052	4.584	4.6e-06	0.0013
BNC2_amp	p53	103	0.333	4.571	4.9e-06	0.0013
MAP2K4_amp	Hypoxia	43	0.927	4.570	5e-06	0.0013
WHSC1L1_amp	p53	438	0.163	4.496	7e-06	0.0018
CDK4_del	Trail	2	0.144	4.468	8e-06	0.002
GNAS_amp	MAPK	281	0.269	4.421	1e-05	0.0024
SMARCA4_amp	PI3K	168	0.619	4.378	1.2e-05	0.0029
FGFR1_amp	p53	427	0.158	4.313	1.6e-05	0.0038
ARHGEF2_amp	JAK-STAT	437	-0.300	-4.280	1.9e-05	0.0043
MNDA_amp	JAK-STAT	368	-0.323	-4.271	2e-05	0.0044
SPRR3_amp	JAK-STAT	430	-0.301	-4.262	2.1e-05	0.0044
SPOP_del	EGFR	9	0.141	4.243	2.2e-05	0.0047
CCND1_amp	p53	616	0.131	4.237	2.3e-05	0.0047
CARM1_amp	PI3K	162	0.605	4.207	2.6e-05	0.0052
FCRL4_amp	JAK-STAT	366	-0.319	-4.206	2.6e-05	0.0052
PRRX1_amp	JAK-STAT	337	-0.332	-4.202	2.7e-05	0.0052
CTTN_amp	PI3K	564	0.331	4.142	3.5e-05	0.0066
HDAC3_del	EGFR	17	0.104	4.137	3.6e-05	0.0066
KEAP1_amp	PI3K	156	0.601	4.099	4.2e-05	0.0076
FXR1_amp	p53	703	0.119	4.080	4.6e-05	0.0081
WT1_amp	p53	89	0.318	4.057	5e-05	0.0087
RB1_del	JAK-STAT	367	-0.310	-4.038	5.5e-05	0.0093
MAP4K1_amp	PI3K	259	0.465	4.034	5.6e-05	0.0093
MUC20_amp	p53	645	0.122	4.021	5.9e-05	0.0096
SETDB1_amp	JAK-STAT	511	-0.261	-4.010	6.1e-05	0.0098
CDC73_amp	JAK-STAT	279	-0.347	-4.009	6.2e-05	0.0098
THRAP3_del	p53	14	0.790	3.993	6.6e-05	0.01
DHX9_amp	JAK-STAT	297	-0.335	-3.987	6.8e-05	0.01
ARAP3_del	EGFR	18	0.097	3.983	6.9e-05	0.01
HLF_amp	JAK-STAT	192	-0.405	-3.957	7.7e-05	0.011
ABL2_amp	TNFa	296	-0.022	-3.946	8e-05	0.012
MYC_amp	p53	971	0.102	3.924	8.8e-05	0.013
RB1_del	p53	367	0.153	3.906	9.5e-05	0.013
DDX5_amp	MAPK	251	-0.259	-3.866	0.00011	0.015
DLG1_amp	p53	652	0.117	3.862	0.00011	0.015
PTPRF_del	p53	11	0.863	3.869	0.00011	0.015
CTTN_amp	p53	564	0.123	3.826	0.00013	0.017
PSIP1_amp	p53	92	0.295	3.829	0.00013	0.017
CDH1_del	EGFR	84	0.042	3.803	0.00014	0.018
PPM1D_amp	MAPK	263	-0.250	-3.819	0.00014	0.017
TP53_amp	MAPK	19	0.906	3.810	0.00014	0.018
ASXL1_del	Trail	1	0.172	3.793	0.00015	0.019
PSME3_amp	TGFb	57	-0.740	-3.777	0.00016	0.02
BNC2_amp	PI3K	103	0.676	3.747	0.00018	0.022
TP53BP1_amp	Hypoxia	13	1.379	3.731	0.00019	0.023
ARID4B_amp	TNFa	346	-0.019	-3.722	0.0002	0.023
RPSAP58_amp	PI3K	93	0.701	3.706	0.00021	0.025
DHX9_amp	TNFa	297	-0.020	-3.687	0.00023	0.026
ASPM_amp	JAK-STAT	279	-0.318	-3.677	0.00024	0.027
PIP5K1A_amp	JAK-STAT	502	-0.240	-3.661	0.00025	0.028
MYC_amp	TNFa	971	0.012	3.657	0.00026	0.028
MSR1_del	MAPK	389	0.194	3.640	0.00028	0.03
NCF2_amp	JAK-STAT	306	-0.300	-3.620	0.0003	0.032
NF1_amp	p53	102	0.266	3.617	0.0003	0.032
PSIP1_amp	PI3K	92	0.689	3.611	0.00031	0.032
CBFB_del	EGFR	56	0.048	3.604	0.00032	0.033
NCF2_amp	TNFa	306	-0.019	-3.593	0.00033	0.034
CLTC_amp	JAK-STAT	251	-0.322	-3.585	0.00034	0.034
PIK3CA_amp	p53	732	0.103	3.579	0.00035	0.035
MED24_amp	JAK-STAT	261	-0.314	-3.568	0.00036	0.036
PIK3R3_del	p53	14	0.707	3.573	0.00036	0.035
CDK12_amp	JAK-STAT	310	-0.287	-3.541	0.0004	0.038
CIC_amp	Hypoxia	101	0.469	3.543	0.0004	0.038
DDX5_amp	TGFb	251	-0.336	-3.542	0.0004	0.038

PLCG1_amp	MAPK	148	0.291	3.539	0.0004	0.038
CSNK2A1_amp	VEGF	90	0.023	3.538	0.00041	0.038
LCP1_del	TGFb	186	-0.394	-3.538	0.00041	0.038
TBL1XR1_amp	p53	708	0.103	3.518	0.00044	0.04
SPOP_amp	JAK-STAT	168	-0.384	-3.514	0.00045	0.04
CUL3_del	PI3K	74	0.749	3.507	0.00046	0.041
EPHA4_del	PI3K	61	0.820	3.493	0.00048	0.043
CLTC_amp	MAPK	251	-0.233	-3.488	0.00049	0.043
DIS3_del	Hypoxia	123	0.414	3.486	0.00049	0.043
CEBPA_amp	p53	254	0.163	3.475	0.00051	0.044
MECOM_amp	p53	781	0.097	3.470	0.00052	0.044
DDX5_amp	JAK-STAT	251	-0.312	-3.468	0.00053	0.044
PPM1D_amp	JAK-STAT	263	-0.305	-3.468	0.00053	0.044
WNK1_del	VEGF	31	0.036	3.457	0.00055	0.046
ERBB2_amp	JAK-STAT	378	-0.254	-3.453	0.00056	0.046
CSNK2A1_amp	TGFb	90	0.527	3.447	0.00057	0.047
AXIN2_amp	TGFb	238	-0.335	-3.438	0.00059	0.048
CREBBP_del	p53	53	0.346	3.432	0.0006	0.048
EIF4G1_amp	p53	664	0.103	3.433	0.0006	0.048

Table B12: Perturbation-response genes vs. CNAs (pan-cancer)

CNA	Pathway	Size	Effect	Wald stat.	P-value	FDR
CDKN2A_del	EGFR	1060	0.744	22.891	1.6e-111	1.6e-107
CDKN2A_del	MAPK	1060	0.683	20.849	1.9e-93	9.6e-90
CAPN7_del	Hypoxia	72	1.992	20.656	1.1e-91	3.6e-88
CLASP2_del	Hypoxia	76	1.914	20.422	1.1e-89	2.9e-86
FXR1_amp	p53	705	0.740	20.244	2.3e-88	4.6e-85
PIK3CA_amp	p53	735	0.717	19.938	7.6e-86	1.3e-82
MLH1_del	Hypoxia	83	1.775	19.754	3.5e-84	5e-81
EIF4G1_amp	p53	665	0.740	19.727	4e-84	5e-81
TBL1XR1_amp	p53	710	0.716	19.639	2.1e-83	2.3e-80
MYD88_del	Hypoxia	76	1.834	19.516	2.9e-82	2.9e-79
TGFBR2_del	Hypoxia	105	1.557	19.381	3.2e-81	2.9e-78
VHL_del	Hypoxia	79	1.780	19.231	4.5e-80	3.7e-77
RFC4_amp	p53	637	0.730	19.119	2.9e-79	2.2e-76
EIF4A2_amp	p53	639	0.728	19.097	4.3e-79	3e-76
ITGA9_del	Hypoxia	84	1.705	19.002	3.3e-78	2.2e-75
PIK3CA_amp	NFkB	735	0.720	18.844	4e-77	2.5e-74
MECOM_amp	NFkB	783	0.699	18.773	1.5e-76	8.5e-74
DLG1_amp	NFkB	654	0.754	18.720	3.6e-76	2e-73
DLG1_amp	p53	654	0.705	18.641	1.4e-75	7.6e-73
TBL1XR1_amp	NFkB	710	0.721	18.619	2.2e-75	1.1e-72
FXR1_amp	NFkB	705	0.719	18.457	3.8e-74	1.8e-71
MUC20_amp	NFkB	647	0.746	18.403	9.8e-74	4.4e-71
PIK3CA_amp	TNFa	735	0.687	18.291	7.1e-73	3e-70
HDAC3_amp	Hypoxia	99	1.662	18.238	9.5e-73	3.9e-70
MECOM_amp	TNFa	783	0.664	18.147	8.6e-72	3.4e-69
ARAP3_amp	Hypoxia	100	1.641	18.092	1.2e-71	4.7e-69
DLG1_amp	TNFa	654	0.717	18.116	1.4e-71	5.3e-69
TBL1XR1_amp	TNFa	710	0.688	18.051	4.5e-71	1.6e-68
EIF4G1_amp	NFkB	665	0.720	18.010	9.2e-71	3.1e-68
MUC20_amp	p53	647	0.686	17.985	1.4e-70	4.6e-68
FXR1_amp	TNFa	705	0.686	17.918	4.4e-70	1.4e-67
EIF4A2_amp	NFkB	639	0.728	17.869	1e-69	3.2e-67
SRGAP3_del	Hypoxia	93	1.529	17.827	2.4e-69	7.3e-67
RFC4_amp	NFkB	637	0.726	17.787	4.1e-69	1.2e-66
MUC20_amp	TNFa	647	0.707	17.754	7.1e-69	2e-66
EIF4G1_amp	TNFa	665	0.690	17.558	2e-67	5.5e-65
EIF4A2_amp	TNFa	639	0.702	17.540	2.7e-67	7.2e-65
RFC4_amp	TNFa	637	0.700	17.460	1e-66	2.7e-64
NPM1_amp	Hypoxia	143	1.315	17.238	2.6e-65	6.6e-63
CTNNB1_del	Hypoxia	91	1.493	17.247	4.6e-65	1.1e-62
ACSL6_amp	Hypoxia	99	1.573	17.170	7.5e-65	1.8e-62
ARHGAP26_amp	Hypoxia	118	1.436	17.149	1.1e-64	2.6e-62
G3BP1_amp	Hypoxia	103	1.523	16.993	1.4e-63	3.3e-61
FAT2_amp	Hypoxia	105	1.491	16.777	4.8e-62	1.1e-59
CDKN2A_del	p53	1060	0.540	16.752	1.1e-61	2.5e-59
SETD2_del	Hypoxia	120	1.264	16.624	1.2e-60	2.6e-58
AFF4_amp	Hypoxia	103	1.486	16.504	3.9e-60	8.2e-58
MECOM_amp	p53	783	0.579	16.367	5.3e-59	1.1e-56
CSNK1G3_amp	Hypoxia	86	1.601	16.214	3.8e-58	7.6e-56
CDKN2A_del	TNFa	1060	0.526	16.157	1.4e-57	2.7e-55

PIK3CA_amp	Hypoxia	735	0.618	15.893	8.3e-56	1.6e-53
FXR1_amp	MAPK	705	0.621	15.878	1e-55	2e-53
TBL1XR1_amp	Hypoxia	710	0.626	15.868	1.2e-55	2.3e-53
FXR1_amp	Hypoxia	705	0.627	15.852	1.5e-55	2.9e-53
RBM5_del	Hypoxia	121	1.187	15.670	3.2e-54	5.8e-52
NSD1_amp	Hypoxia	178	1.075	15.622	3.8e-54	6.8e-52
PIK3CA_amp	MAPK	735	0.599	15.588	8.6e-54	1.5e-51
RHOA_del	Hypoxia	124	1.154	15.425	1.3e-52	2.2e-50
EIF4G1_amp	MAPK	665	0.619	15.402	1.4e-52	2.4e-50
TBL1XR1_amp	MAPK	710	0.597	15.286	7.9e-52	1.3e-49
EGFR_amp	MAPK	351	0.815	15.277	1.1e-51	1.9e-49
MECOM_amp	MAPK	783	0.570	15.229	1.8e-51	2.9e-49
BAP1_del	Hypoxia	137	1.089	15.243	1.9e-51	3e-49
EIF4G1_amp	Hypoxia	665	0.603	14.839	5.4e-49	8.4e-47
PBRM1_del	Hypoxia	133	1.074	14.838	6.8e-49	1e-46
DLG1_amp	MAPK	654	0.600	14.784	1.2e-48	1.8e-46
EIF4A2_amp	MAPK	639	0.605	14.766	1.5e-48	2.3e-46
RFC4_amp	MAPK	637	0.601	14.668	6.3e-48	9.2e-46
MECOM_amp	Hypoxia	783	0.545	14.337	6.7e-46	9.6e-44
WNT5A_del	Hypoxia	107	1.163	14.336	8.1e-46	1.2e-43
MUC20_amp	MAPK	647	0.585	14.314	9e-46	1.3e-43
DLG1_amp	Hypoxia	654	0.585	14.286	1.3e-45	1.8e-43
RASA2_amp	p53	314	0.757	14.176	6.2e-45	8.4e-43
XRN1_amp	p53	330	0.737	14.119	1.3e-44	1.8e-42
ATR_amp	p53	330	0.735	14.072	2.6e-44	3.4e-42
TRIO_amp	TNFa	438	0.684	14.039	4.2e-44	5.5e-42
TFDP2_amp	p53	328	0.730	13.932	1.7e-43	2.2e-41
DLG1_amp	EGFR	654	0.557	13.894	2.9e-43	3.7e-41
EIF4A2_amp	Hypoxia	639	0.574	13.883	3.4e-43	4.3e-41
RFC4_amp	Hypoxia	637	0.573	13.850	5.3e-43	6.6e-41
TRIO_amp	NFkB	438	0.678	13.777	1.5e-42	1.8e-40
MUC20_amp	Hypoxia	647	0.568	13.769	1.6e-42	1.9e-40
MUC20_amp	EGFR	647	0.551	13.658	6.9e-42	8.3e-40
CCT5_amp	TNFa	406	0.688	13.601	1.5e-41	1.8e-39
EIF4A2_amp	EGFR	639	0.549	13.514	4.7e-41	5.5e-39
EIF4G1_amp	EGFR	665	0.537	13.451	1.1e-40	1.3e-38
CTTN_amp	p53	564	0.563	13.383	2.2e-40	2.6e-38
RFC4_amp	EGFR	637	0.545	13.391	2.4e-40	2.7e-38
CCT5_amp	NFkB	406	0.684	13.366	3.4e-40	3.8e-38
CDKN2A_del	NFkB	1060	0.449	13.338	4.8e-40	5.3e-38
SMAD4_del	EGFR	208	0.902	13.279	1.2e-39	1.3e-37
FXR1_amp	EGFR	705	0.515	13.237	1.8e-39	1.9e-37
CTTN_amp	TNFa	564	0.555	13.117	7.1e-39	7.6e-37
CTTN_amp	MAPK	564	0.570	13.110	7.8e-39	8.3e-37
CDKN2A_del	Hypoxia	1060	0.426	13.090	1.2e-38	1.2e-36
PIK3CA_amp	EGFR	735	0.498	13.024	2.8e-38	2.9e-36
KRAS_amp	MAPK	277	0.793	12.983	4.4e-38	4.5e-36
MECOM_amp	JAK-STAT	783	0.493	12.963	6e-38	6.1e-36
DLG1_amp	JAK-STAT	654	0.533	12.948	7.2e-38	7.2e-36
MUC20_amp	JAK-STAT	647	0.534	12.894	1.4e-37	1.4e-35
TBL1XR1_amp	EGFR	710	0.497	12.783	5.8e-37	5.7e-35
PIK3CA_amp	JAK-STAT	735	0.497	12.739	1e-36	9.7e-35
ATR_amp	NFkB	330	0.709	12.701	1.6e-36	1.5e-34
TBL1XR1_amp	JAK-STAT	710	0.503	12.702	1.6e-36	1.5e-34
CTTN_amp	EGFR	564	0.545	12.685	1.7e-36	1.6e-34
APC_amp	Hypoxia	75	1.355	12.682	1.7e-36	1.6e-34
XRN1_amp	NFkB	330	0.707	12.664	2.6e-36	2.4e-34
MECOM_amp	EGFR	783	0.466	12.535	1.3e-35	1.2e-33
RASA2_amp	NFkB	314	0.709	12.413	5.6e-35	5.2e-33
STAG1_amp	p53	258	0.729	12.410	5.8e-35	5.3e-33
TFDP2_amp	NFkB	328	0.694	12.402	6.4e-35	5.7e-33
NCK1_amp	p53	258	0.726	12.352	1.2e-34	1e-32
FXR1_amp	JAK-STAT	705	0.490	12.332	1.5e-34	1.3e-32
CCND1_amp	TNFa	616	0.493	12.110	1.9e-33	1.7e-31
CTTN_amp	JAK-STAT	564	0.533	12.107	2e-33	1.7e-31
CTTN_amp	NFkB	564	0.526	12.106	2e-33	1.7e-31
XRN1_amp	TNFa	330	0.664	12.083	3e-33	2.5e-31
ATR_amp	TNFa	330	0.663	12.054	4.2e-33	3.6e-31
CTTN_amp	VEGF	564	0.528	12.021	5.6e-33	4.6e-31
CCND1_amp	VEGF	616	0.505	11.980	9.1e-33	7.5e-31
CCND1_amp	p53	616	0.483	11.916	1.9e-32	1.6e-30
CDKN2A_del	PI3K	1060	0.400	11.915	2.2e-32	1.8e-30
TFDP2_amp	TNFa	328	0.653	11.861	4.1e-32	3.3e-30
PIK3CB_amp	p53	293	0.650	11.730	1.9e-31	1.5e-29
CCND1_amp	EGFR	616	0.482	11.681	3e-31	2.4e-29

MYC_amp	JAK-STAT	974	0.402	11.691	3.4e-31	2.7e-29
RASA2_amp	TNFa	314	0.654	11.615	7.1e-31	5.6e-29
CCND1_amp	MAPK	616	0.483	11.535	1.6e-30	1.2e-28
EIF4A2_amp	JAK-STAT	639	0.480	11.539	1.7e-30	1.3e-28
RFC4_amp	JAK-STAT	637	0.478	11.479	3.3e-30	2.5e-28
EIF4G1_amp	JAK-STAT	665	0.467	11.435	5.5e-30	4.2e-28
CCND1_amp	JAK-STAT	616	0.475	11.214	6e-29	4.5e-27
PIK3CB_amp	NFkB	293	0.663	11.217	6.3e-29	4.7e-27
TRIO_amp	MAPK	438	0.544	11.030	5e-28	3.7e-26
ASXL1_amp	PI3K	246	0.665	11.029	5.5e-28	4e-26
EGFR_amp	EGFR	351	0.595	10.988	8.3e-28	6.1e-26
STAG1_amp	NFkB	258	0.683	10.832	4.2e-27	3e-25
NCK1_amp	NFkB	258	0.682	10.820	4.7e-27	3.4e-25
NDRG1_amp	JAK-STAT	744	0.413	10.734	1.3e-26	9.4e-25
TRIO_amp	EGFR	438	0.531	10.729	1.3e-26	9.1e-25
PIK3CB_amp	TNFa	293	0.620	10.659	2.6e-26	1.9e-24
XRN1_amp	Hypoxia	330	0.599	10.604	4.7e-26	3.3e-24
MYC_amp	MAPK	974	0.358	10.505	1.5e-25	1e-23
CCND1_amp	NFkB	616	0.435	10.400	3.7e-25	2.6e-23
TFDP2_amp	Hypoxia	328	0.589	10.393	4.2e-25	2.9e-23
MYC_amp	TNFa	974	0.354	10.389	4.9e-25	3.3e-23
CCT5_amp	MAPK	406	0.529	10.343	7.2e-25	4.9e-23
FRG1B_amp	PI3K	239	0.633	10.346	7.4e-25	5e-23
CCT5_amp	EGFR	406	0.529	10.316	9.5e-25	6.3e-23
ATR_amp	Hypoxia	330	0.582	10.291	1.2e-24	8e-23

B.3 *Pathway scores and drugs*

Pan-cancer

Table B13: Gene Ontology vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
Trametinib	Trail	713	-1.099	-5.824	8.8e-09	2.6e-05
Bleomycin (50 uM)	TNFa	746	-2.292	-5.592	3.2e-08	4.7e-05
RDEA119 (rescreen)	Trail	698	-0.706	-5.459	6.8e-08	6.6e-05
BMN-673	p53	727	-1.321	-4.989	7.7e-07	0.00054
RDEA119	Trail	668	-0.743	-4.957	9.2e-07	0.00054
MP470	JAK-STAT	730	1.254	4.823	1.7e-06	0.00084
Bleomycin (50 uM)	NFkB	746	-1.242	-4.621	4.5e-06	0.0019
Nutlin-3a	p53	676	-0.919	-4.572	5.8e-06	0.0021
Bleomycin (50 uM)	JAK-STAT	746	-1.349	-4.317	1.8e-05	0.0059
TAK-715	Hypoxia	735	1.279	4.289	2e-05	0.006
TAK-715	TGFb	735	1.165	4.227	2.7e-05	0.0071
Bleomycin	TNFa	682	-2.314	-4.185	3.2e-05	0.0079
Bleomycin	p53	682	-1.533	-4.016	6.6e-05	0.015
CI-1040	Trail	669	-0.515	-3.954	8.6e-05	0.018
Lenalidomide	NFkB	677	-0.437	-3.890	0.00011	0.022
PD-0325901	Trail	669	-0.586	-3.847	0.00013	0.024
MP470	TGFb	730	1.581	3.753	0.00019	0.026
GSK690693	JAK-STAT	732	0.948	3.755	0.00019	0.026
JNK Inhibitor VIII	NFkB	674	-0.453	-3.756	0.00019	0.026
RO-3306	NFkB	676	-0.595	-3.751	0.00019	0.026
XAV 939	TNFa	730	-0.747	-3.742	0.0002	0.026
BMN-673	NFkB	727	-0.947	-3.739	0.0002	0.026
SN-38	TNFa	741	-1.308	-3.716	0.00022	0.027
Bleomycin (50 uM)	p53	746	-1.056	-3.718	0.00022	0.027
MK-2206	JAK-STAT	658	0.933	3.680	0.00025	0.028
CCT018159	Trail	715	-0.374	-3.677	0.00025	0.028
JNK Inhibitor VIII	TNFa	674	-0.694	-3.671	0.00026	0.028
RDEA119 (rescreen)	TGFb	698	-1.396	-3.645	0.00029	0.03
FTI-277	TNFa	695	-0.654	-3.564	0.00039	0.038
BMN-673	TNFa	727	-1.382	-3.560	0.0004	0.038
Dasatinib	TGFb	287	-3.603	-3.564	0.00043	0.038
YK 4-279	TNFa	627	-1.115	-3.543	0.00043	0.038
EHT 1864	JAK-STAT	737	0.525	3.536	0.00043	0.038
Temozolomide	p53	721	-0.377	-3.513	0.00047	0.04
Afatinib (rescreen)	Trail	727	-0.524	-3.499	0.0005	0.041
PD-0332991	Trail	659	-0.509	-3.448	0.0006	0.049
AUY922	TNFa	681	-1.299	-3.435	0.00063	0.049
Trametinib	TGFb	713	-1.930	-3.429	0.00064	0.049

FTI-277	PI3K	695	0.616	3.427	0.00065	0.049

Table B14: Reactome vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
RDEA119 (rescreen)	Trail	698	-0.973	-6.328	4.6e-10	6.9e-07
Trametinib	Trail	713	-1.418	-6.320	4.7e-10	6.9e-07
RDEA119	TNFa	668	-1.252	-5.127	3.9e-07	0.00038
RDEA119	Trail	668	-0.899	-4.981	8.2e-07	0.0006
Docetaxel	TGFb	676	-1.453	-4.797	2e-06	0.0012
TG101348	Hypoxia	734	-1.058	-4.749	2.5e-06	0.0012
Bleomycin (50 uM)	TGFb	746	-1.792	-4.662	3.7e-06	0.0014
Trametinib	TNFa	713	-1.467	-4.667	3.7e-06	0.0014
MP470	VEGF	730	1.899	4.626	4.4e-06	0.0014
Trametinib	VEGF	713	-2.450	-4.527	7.1e-06	0.0021
MP470	JAK-STAT	730	1.617	4.315	1.8e-05	0.0048
MP470	EGFR	730	2.594	4.290	2e-05	0.0049
Afatinib (rescreen)	Trail	727	-0.764	-4.274	2.2e-05	0.0049
Trametinib	TGFb	713	-1.791	-4.154	3.7e-05	0.0075
AZD6244	TNFa	658	-1.023	-4.129	4.1e-05	0.0075
AZD6244	TNFa	658	-1.023	-4.129	4.1e-05	0.0075
RDEA119 (rescreen)	VEGF	698	-1.518	-4.086	4.9e-05	0.0085
Trametinib	EGFR	713	-3.256	-4.037	6e-05	0.0093
EHT 1864	JAK-STAT	737	0.846	4.029	6.2e-05	0.0093
Cetuximab	JAK-STAT	688	-1.013	-4.024	6.4e-05	0.0093
TL-2-105	NFkB	735	0.777	3.994	7.2e-05	0.01
Bleomycin (50 uM)	TNFa	746	-1.106	-3.950	8.6e-05	0.011
Afatinib (rescreen)	NFkB	727	-1.128	-3.945	8.8e-05	0.011
Pyrimethamine	TGFb	292	-2.045	-3.904	0.00012	0.014
EHT 1864	p53	737	-0.723	-3.868	0.00012	0.014
Cetuximab	Trail	688	-0.452	-3.867	0.00012	0.014
RDEA119	VEGF	668	-1.629	-3.816	0.00015	0.016
PD-0325901	Trail	669	-0.699	-3.814	0.00015	0.016
NSC-207895	TGFb	730	-1.077	-3.789	0.00016	0.017
RDEA119 (rescreen)	TNFa	698	-0.815	-3.762	0.00018	0.017
XAV 939	NFkB	730	-0.597	-3.760	0.00018	0.017
(5Z)-7-Oxozeaenol	VEGF	729	-1.267	-3.743	0.0002	0.018
Bleomycin (50 uM)	JAK-STAT	746	-1.639	-3.733	0.0002	0.018
Obatoclax Mesylate	TGFb	684	-1.373	-3.723	0.00021	0.018
A-770041	PI3K	289	-4.074	-3.742	0.00022	0.018
SN-38	TGFb	741	-1.211	-3.701	0.00023	0.018
Nutlin-3a	Trail	676	-0.541	-3.705	0.00023	0.018
Docetaxel	NFkB	676	-0.954	-3.706	0.00023	0.018
YK 4-279	TGFb	627	-1.066	-3.700	0.00024	0.018
BMN-673	TGFb	727	-1.338	-3.677	0.00025	0.018
RDEA119 (rescreen)	EGFR	698	-2.032	-3.677	0.00025	0.018
PFI-1	TGFb	744	-0.786	-3.652	0.00028	0.019
RDEA119	JAK-STAT	668	-1.430	-3.654	0.00028	0.019
CI-1040	Trail	669	-0.569	-3.643	0.00029	0.019
RDEA119 (rescreen)	TGFb	698	-1.069	-3.627	0.00031	0.02
FMK	Hypoxia	623	-0.504	-3.599	0.00035	0.022
XL-880	Hypoxia	730	-0.799	-3.587	0.00036	0.022
Cytarabine	Hypoxia	673	-1.034	-3.579	0.00037	0.022
RDEA119	TGFb	668	-1.224	-3.576	0.00038	0.022
Dabrafenib	Trail	692	-0.663	-3.570	0.00038	0.022
XL-880	TGFb	730	-0.975	-3.558	0.0004	0.023
AC220	JAK-STAT	734	0.763	3.526	0.00045	0.025
MLN4924	TGFb	553	-1.276	-3.523	0.00046	0.025
Cetuximab	VEGF	688	-0.984	-3.520	0.00046	0.025
RDEA119	EGFR	668	-2.224	-3.524	0.00046	0.025
Bleomycin (50 uM)	VEGF	746	-1.691	-3.478	0.00054	0.027
(5Z)-7-Oxozeaenol	TNFa	729	-0.682	-3.473	0.00055	0.027
Axitinib	JAK-STAT	673	0.997	3.474	0.00055	0.027
PD-0325901	TNFa	669	-0.860	-3.448	0.0006	0.03
ABT-888	p53	676	-0.544	-3.435	0.00063	0.03
Camptothecin	TGFb	675	-1.242	-3.433	0.00064	0.03
Bleomycin (50 uM)	MAPK	746	-1.318	-3.423	0.00065	0.031
NPK76-II-72-1	TGFb	733	-1.012	-3.409	0.00069	0.031
SB52334	TNFa	732	0.631	3.406	0.0007	0.031
XAV 939	JAK-STAT	730	-0.721	-3.404	0.0007	0.031
Docetaxel	JAK-STAT	676	-1.175	-3.403	0.00071	0.031
TAK-715	EGFR	735	1.342	3.375	0.00078	0.034
AZD7762	Hypoxia	676	-0.847	-3.367	0.00081	0.035

Afatinib	NFkB	675	-0.903	-3.355	0.00084	0.035
MP470	MAPK	730	1.083	3.331	0.00091	0.038
JW-7-24-1	Hypoxia	734	-0.748	-3.321	0.00094	0.039
Cytarabine	TGFb	673	-1.149	-3.297	0.001	0.041
MK-2206	MAPK	658	1.047	3.301	0.001	0.041
Bleomycin	TGFb	682	-1.677	-3.265	0.0012	0.044
VX-11e	TGFb	733	-0.924	-3.244	0.0012	0.046
TW 37	TGFb	735	-0.793	-3.260	0.0012	0.044
Trametinib	JAK-STAT	713	-1.608	-3.263	0.0012	0.044
SN-38	JAK-STAT	741	-1.216	-3.262	0.0012	0.044
XAV 939	TGFb	730	-0.604	-3.223	0.0013	0.048
Nutlin-3a	p53	676	-0.921	-3.240	0.0013	0.046
Phenformin	JAK-STAT	728	1.306	3.220	0.0013	0.048

Table B15: SPIA vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
RDEA119 (rescreen)	Trail	678	-0.013	-5.536	4.5e-08	7.7e-05
Trametinib	Trail	692	-0.019	-5.445	7.3e-08	7.7e-05
Bleomycin (50 uM)	Trail	724	-0.016	-5.206	2.5e-07	0.00016
Bleomycin (50 uM)	MAPK	724	-0.019	-5.134	3.7e-07	0.00016
RDEA119	Trail	648	-0.014	-5.135	3.8e-07	0.00016
Cetuximab	Trail	668	-0.009	-5.093	4.6e-07	0.00016
NVP-BEZ235	MAPK	650	-0.012	-4.703	3.2e-06	0.00096
PI-103	JAK-STAT	707	0.089	4.279	2.2e-05	0.0051
XAV 939	MAPK	708	-0.007	-4.275	2.2e-05	0.0051
PLX4720 (rescreen)	Trail	708	-0.008	-4.181	3.3e-05	0.007
PD-0325901	Trail	649	-0.012	-4.145	3.9e-05	0.0075
XAV 939	Trail	708	-0.006	-4.063	5.4e-05	0.0096
NG-25	JAK-STAT	713	0.054	3.911	0.0001	0.016
GSK2126458	JAK-STAT	714	0.063	3.900	0.00011	0.016
SN-38	Trail	719	-0.010	-3.832	0.00014	0.02
NVP-BEZ235	Trail	650	-0.009	-3.808	0.00015	0.02
XAV 939	NFkB	708	-0.007	-3.751	0.00019	0.021
Temozolomide	Trail	700	-0.004	-3.736	0.0002	0.021
(5Z)-7-Oxozeaenol	Trail	708	-0.008	-3.713	0.00022	0.021
Axitinib	EGFR	654	0.010	3.718	0.00022	0.021
AZD6482	NFkB	670	-0.010	-3.720	0.00022	0.021
AZD6482	NFkB	670	-0.010	-3.720	0.00022	0.021
RDEA119 (rescreen)	JAK-STAT	678	-0.050	-3.701	0.00023	0.021
Nutlin-3a	Trail	656	-0.008	-3.686	0.00025	0.022
Afatinib (rescreen)	Trail	705	-0.010	-3.669	0.00026	0.022
CI-1040	Trail	650	-0.009	-3.664	0.00027	0.022
SN-38	MAPK	719	-0.011	-3.616	0.00032	0.025
CEP-701	MAPK	657	-0.010	-3.558	0.0004	0.03
Afatinib	EGFR	656	-0.012	-3.516	0.00047	0.034
TL-1-85	JAK-STAT	713	0.044	3.484	0.00053	0.037
Camptothecin	MAPK	656	-0.012	-3.459	0.00058	0.04
CMK	EGFR	279	0.017	3.436	0.00069	0.046
Afatinib (rescreen)	EGFR	705	-0.012	-3.389	0.00074	0.047
Bleomycin	Trail	661	-0.015	-3.386	0.00075	0.047
Methotrexate	VEGF	656	-0.019	-3.372	0.00079	0.048
CAL-101	NFkB	714	-0.007	-3.361	0.00082	0.048
Pyrimethamine	VEGF	285	-0.031	-3.379	0.00084	0.048
Axitinib	NFkB	654	0.008	3.344	0.00088	0.049

Table B16: Pathifier vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
HG-6-64-1	VEGF	642	-1.375	-4.292	2e-05	0.031
Temsirolimus	VEGF	622	-1.331	-4.283	2.1e-05	0.031
Vorinostat	JAK-STAT	629	0.848	3.881	0.00011	0.084
FK866	JAK-STAT	668	2.121	3.900	0.00011	0.084
piperlongumine	JAK-STAT	688	0.696	3.811	0.00015	0.084
AZ628	VEGF	271	-1.896	-3.812	0.00017	0.084
Ruxolitinib	JAK-STAT	683	0.583	3.648	0.00028	0.099
AICAR	p53	627	-0.814	-3.611	0.00033	0.099
CP724714	JAK-STAT	686	0.682	3.604	0.00034	0.099
VX-702	JAK-STAT	626	0.590	3.602	0.00034	0.099

Table B17: PARADIGM vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
ABT-869	Trail	732	4.197	3.528	0.00045	0.38
XMD13-2	TNFa	671	-2.326	-3.368	0.0008	0.38
17-AAG	PI3K	447	0.123	3.307	0.001	0.38
Z-LLNle-CHO	VEGF	43	5.884	3.289	0.0011	0.38
GSK-650394	TNFa	622	-3.478	-3.246	0.0012	0.38
CCT018159	PI3K	476	0.074	3.228	0.0013	0.38
Ruxolitinib	Trail	731	3.469	3.113	0.0019	0.38
JW-7-24-1	TNFa	672	-2.511	-3.062	0.0023	0.38
XAV 939	MAPK	730	0.124	3.049	0.0024	0.38
JW-7-24-1	JAK-STAT	734	0.182	3.039	0.0025	0.38

Table B18: Perturbation-response genes vs. drugs (pan-cancer)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
Nutlin-3a	p53	676	-0.562	-11.662	1.2e-28	3.5e-25
Trametinib	EGFR	713	-0.847	-8.677	3e-17	4.3e-14
Trametinib	MAPK	713	-0.727	-8.153	1.7e-15	1.7e-12
RDEA119 (rescreen)	EGFR	698	-0.486	-7.201	1.6e-12	1.2e-09
RDEA119	EGFR	668	-0.547	-7.043	4.9e-12	2.9e-09
RDEA119 (rescreen)	MAPK	698	-0.423	-6.826	2e-11	9.5e-09
RDEA119	MAPK	668	-0.474	-6.780	2.7e-11	1.1e-08
Dabrafenib	MAPK	692	-0.460	-6.209	9.4e-10	3.4e-07
(5Z)-7-Oxozeaenol	MAPK	729	-0.343	-6.041	2.5e-09	8.1e-07
VX-11e	MAPK	733	-0.341	-5.703	1.7e-08	5.1e-06
AZ628	MAPK	292	-0.745	-5.781	2.1e-08	5.5e-06
XAV 939	TNFa	730	-0.203	-5.208	2.5e-07	5.6e-05
Dabrafenib	EGFR	692	-0.429	-5.211	2.5e-07	5.6e-05
CCT018159	MAPK	715	-0.233	-4.903	1.2e-06	0.00025
CI-1040	EGFR	669	-0.339	-4.849	1.6e-06	0.00029
Bleomycin (50 uM)	TNFa	746	-0.393	-4.838	1.6e-06	0.00029
CI-1040	MAPK	669	-0.291	-4.685	3.4e-06	0.00059
(5Z)-7-Oxozeaenol	EGFR	729	-0.293	-4.653	3.9e-06	0.00063
PLX4720 (rescreen)	TNFa	730	-0.219	-4.597	5.1e-06	0.00078
PD-0325901	EGFR	669	-0.366	-4.565	6e-06	0.00087
NSC-207895	p53	730	0.242	4.432	1.1e-05	0.0015
CCT018159	EGFR	715	-0.234	-4.400	1.3e-05	0.0017
EHT 1864	JAK-STAT	737	0.161	4.317	1.8e-05	0.0023
PD-0325901	MAPK	669	-0.311	-4.310	1.9e-05	0.0023
AZ628	EGFR	292	-0.638	-4.334	2.1e-05	0.0024
VX-11e	EGFR	733	-0.283	-4.239	2.5e-05	0.0027
XAV 939	TGFb	730	-0.214	-4.244	2.5e-05	0.0027
XAV 939	NFkB	730	-0.166	-4.236	2.6e-05	0.0027
Bleomycin	NFkB	682	-0.461	-4.219	2.8e-05	0.0028
Bleomycin	TNFa	682	-0.462	-4.191	3.2e-05	0.0031
AZD7762	PI3K	676	-0.253	-4.178	3.3e-05	0.0031
Bleomycin (50 uM)	NFkB	746	-0.331	-4.041	5.9e-05	0.0054
Thapsigargin	MAPK	680	-0.382	-4.020	6.5e-05	0.0057
SB590885	TGFb	660	0.254	4.003	7e-05	0.006
Afatinib	Hypoxia	675	0.265	3.993	7.3e-05	0.0061
NVP-BHG712	MAPK	733	-0.245	-3.984	7.5e-05	0.0061
Axitinib	JAK-STAT	673	0.206	3.954	8.5e-05	0.0067
NVP-TAE684	p53	292	0.384	3.980	8.9e-05	0.0067
Afatinib (rescreen)	Hypoxia	727	0.279	3.940	9e-05	0.0067
SN-38	TNFa	741	-0.271	-3.921	9.7e-05	0.007
SB590885	MAPK	660	-0.194	-3.880	0.00012	0.0082
17-AAG	EGFR	677	-0.327	-3.864	0.00012	0.0085
Gefitinib	Hypoxia	673	0.188	3.829	0.00014	0.0096
FTI-277	EGFR	695	-0.153	-3.822	0.00014	0.0096
CAL-101	Trail	734	-0.264	-3.788	0.00016	0.01
Bleomycin (50 uM)	VEGF	746	-0.312	-3.801	0.00016	0.01
TGX221	Hypoxia	290	-0.293	-3.811	0.00017	0.011
Bleomycin (50 uM)	MAPK	746	-0.312	-3.744	0.0002	0.012
Camptothecin	TNFa	675	-0.287	-3.705	0.00023	0.014
SB-505124	JAK-STAT	741	0.135	3.690	0.00024	0.014
FMK	MAPK	623	-0.137	-3.670	0.00026	0.015
Cytarabine	MAPK	673	-0.272	-3.659	0.00027	0.015
CH5424802	MAPK	732	-0.151	-3.655	0.00028	0.015
GSK429286A	p53	734	0.150	3.640	0.00029	0.016
SN-38	NFkB	741	-0.252	-3.631	0.0003	0.016

RDEA119 (rescreen)	TNFa	698	-0.224	-3.611	0.00033	0.017
T0901317	JAK-STAT	728	0.128	3.605	0.00033	0.017
FMK	VEGF	623	-0.128	-3.597	0.00035	0.018
MP470	TGFb	730	0.316	3.581	0.00037	0.018
Camptothecin	NFkB	675	-0.276	-3.570	0.00038	0.019
SB52334	TGFb	732	0.242	3.545	0.00042	0.02
PLX4720 (rescreen)	NFkB	730	-0.168	-3.522	0.00046	0.021
AZD6482	TNFa	691	-0.213	-3.496	0.0005	0.023
AZD6482	TNFa	691	-0.213	-3.496	0.0005	0.023
SN-38	MAPK	741	-0.246	-3.478	0.00054	0.024
Docetaxel	MAPK	676	-0.229	-3.478	0.00054	0.024
Docetaxel	TNFa	676	-0.227	-3.454	0.00059	0.026
FH535	TGFb	686	0.255	3.435	0.00063	0.027
STF-62247	p53	732	0.119	3.431	0.00064	0.027
DMOG	TNFa	692	-0.211	-3.408	0.0007	0.029
AUY922	MAPK	681	-0.257	-3.403	0.00071	0.029
TAK-715	TGFb	735	0.196	3.398	0.00072	0.029
Temsirolimus	VEGF	666	-0.245	-3.381	0.00077	0.031
WH-4-023	TGFb	287	-0.645	-3.374	0.00085	0.033
Gemcitabine	MAPK	685	-0.415	-3.347	0.00086	0.033
FTI-277	MAPK	695	-0.122	-3.338	0.00089	0.033
Temsirolimus	PI3K	666	-0.229	-3.335	0.0009	0.033
(5Z)-7-Oxozeaenol	VEGF	729	-0.191	-3.332	0.00091	0.033
TGX221	TGFb	290	-0.359	-3.352	0.00092	0.033
SNX-2112	MAPK	725	-0.290	-3.327	0.00092	0.033
PF-562271	MAPK	680	-0.167	-3.326	0.00093	0.033
Obatoclax Mesylate	MAPK	684	-0.265	-3.320	0.00095	0.034
CP724714	TGFb	734	0.179	3.317	0.00096	0.034
KIN001-270	p53	733	0.103	3.310	0.00098	0.034
EHT 1864	TGFb	737	0.165	3.302	0.001	0.035
FH535	Hypoxia	686	0.198	3.294	0.001	0.035
Genentech Cpd 10	p53	734	0.211	3.273	0.0011	0.037
PLX4720 (rescreen)	MAPK	730	-0.156	-3.241	0.0012	0.039
BX-795	PI3K	676	-0.176	-3.245	0.0012	0.039
AZD6244	PI3K	658	0.222	3.266	0.0012	0.038
AZD6244	PI3K	658	0.222	3.266	0.0012	0.038
Cetuximab	TNFa	688	-0.152	-3.261	0.0012	0.038
Trametinib	p53	713	-0.270	-3.243	0.0012	0.039
17-AAG	VEGF	677	-0.241	-3.240	0.0013	0.039
SL 0101-1	MAPK	664	-0.113	-3.236	0.0013	0.039
Afatinib (rescreen)	p53	727	-0.207	-3.204	0.0014	0.043
BI-2536	p53	288	0.341	3.213	0.0015	0.044
Thapsigargin	p53	680	0.269	3.190	0.0015	0.044
Bleomycin	MAPK	682	-0.353	-3.196	0.0015	0.044
Cetuximab	NFkB	688	-0.148	-3.178	0.0016	0.045
MP470	JAK-STAT	730	0.212	3.157	0.0017	0.048

Tissue specific

Table B19: Gene Ontology vs. drugs (tissue-specific)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
Nutlin-3a	p53	23	-3.551	-4.741	0.00013	0.12
PD-0325901	Hypoxia	40	6.985	4.209	0.00016	0.045
Cisplatin	VEGF	13	-4.200	-5.731	0.00019	0.059
Axitinib	Hypoxia	51	4.209	3.980	0.00023	0.046
SB-715992	p53	25	-2.794	-4.234	0.00034	0.17
Vismodegib	JAK-STAT	29	1.891	3.845	0.0007	0.55
Trametinib	Hypoxia	48	7.143	3.612	0.00076	0.062
AZD6244	p53	22	-4.676	-3.995	0.00078	0.12
AZD6244	p53	22	-4.676	-3.995	0.00078	0.12
PD-0325901	MAPK	40	9.195	3.646	0.00081	0.062
RDEA119	MAPK	40	9.250	3.622	0.00087	0.062
CI-1040	MAPK	45	8.203	3.448	0.0013	0.074
PD-0325901	JAK-STAT	40	3.822	3.396	0.0016	0.076
RDEA119	Hypoxia	40	5.990	3.353	0.0019	0.076
Trametinib	MAPK	48	9.337	3.216	0.0024	0.086
Trametinib	JAK-STAT	48	4.005	3.132	0.003	0.092
CEP-701	TGFb	45	4.243	3.070	0.0037	0.092
CEP-701	MAPK	45	6.160	3.057	0.0039	0.092
PD-0325901	TGFb	40	5.506	3.083	0.0039	0.092
Tipifarnib	PI3K	48	4.180	3.014	0.0042	0.093
CEP-701	EGFR	45	1.715	2.993	0.0046	0.094

Table B20: Reactome vs. drugs (tissue-specific)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
AT-7519	EGFR	20	12.077	5.151	8e-05	0.02
SB-715992	Hypoxia	25	-3.784	-4.506	0.00018	0.17
Afatinib (rescreen)	Trail	23	-1.469	-4.404	0.00027	0.32
AT-7519	VEGF	20	8.657	4.514	0.00031	0.039
Pazopanib	TNFa	49	-2.685	-3.795	0.00043	0.067
ABT-263	Trail	49	3.645	3.748	0.0005	0.098
Ruxolitinib	NFkB	27	5.059	4.015	0.00051	0.37
Trametinib	p53	48	-5.534	-3.640	0.0007	0.067
PD-0325901	VEGF	40	8.932	3.695	0.00071	0.067
GDC0941	p53	31	9.983	3.783	0.00075	0.37
17-AAG	Hypoxia	42	-4.145	-3.456	0.0013	0.096

Table B21: SPIA vs. drugs (tissue-specific)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
OSI-027	NFkB	22	-0.046	-4.469	0.00026	0.14
OSI-027	MAPK	22	-0.060	-4.382	0.00032	0.14
Ruxolitinib	VEGF	21	-0.137	-4.232	0.0005	0.14
GSK2126458	NFkB	27	-0.054	-3.865	0.00074	0.34
GSK2126458	MAPK	22	-0.083	-3.921	0.00092	0.19
AV-951	JAK-STAT	27	0.092	3.760	0.00096	0.34
Shikonin	NFkB	21	-0.027	-3.888	0.0011	0.19
EKB-569	EGFR	46	0.043	3.468	0.0012	0.14
GSK2126458	NFkB	22	-0.062	-3.774	0.0013	0.19
GSK2126458	JAK-STAT	27	0.316	3.595	0.0015	0.35
ZSTK474	MAPK	47	0.037	3.201	0.0025	0.08
Afatinib (rescreen)	EGFR	46	-0.050	-3.203	0.0026	0.08
Afatinib (rescreen)	VEGF	46	0.075	2.987	0.0046	0.08
CH5424802	NFkB	46	-0.023	-2.904	0.0058	0.08
GSK2126458	MAPK	46	0.038	2.849	0.0067	0.08
MP470	PI3K	46	0.170	2.852	0.0067	0.08
Afatinib	EGFR	44	-0.040	-2.719	0.0096	0.098

Table B22: Pathifier vs. drugs (tissue-specific)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
ABT-263	EGFR	47	6.245	5.720	7.6e-07	0.00015
ABT-263	VEGF	46	5.175	4.790	1.8e-05	0.0018
ABT-263	PI3K	47	6.203	4.566	3.7e-05	0.0025
GSK2126458	MAPK	21	5.347	5.096	6.4e-05	0.076
ABT-263	JAK-STAT	45	8.539	4.319	8.3e-05	0.0041
Trametinib	PI3K	47	7.220	4.244	0.00011	0.022
Camptothecin	EGFR	9	4.776	7.713	0.00011	0.046
AT-7519	NFkB	51	6.730	4.145	0.00014	0.0054
Trametinib	VEGF	47	6.736	4.130	0.00016	0.022
5-Fluorouracil	JAK-STAT	27	4.946	4.467	0.00016	0.081
Crizotinib	PI3K	23	3.265	4.514	0.00019	0.081
ABT-263	Trail	48	6.438	3.997	0.00023	0.0076
CH5424802	PI3K	26	4.501	4.295	0.00025	0.081
Trametinib	EGFR	47	6.869	3.901	0.00032	0.025
ZSTK474	VEGF	46	4.391	3.872	0.00035	0.015
PD-0325901	MAPK	40	6.398	3.891	0.0004	0.025
GSK2126458	VEGF	45	4.905	3.793	0.00046	0.015
GSK2126458	TGFb	45	4.479	3.755	0.00052	0.015
Ruxolitinib	MAPK	26	5.113	4.006	0.00052	0.097
Gemcitabine	p53	25	7.826	4.004	0.00056	0.097
AZD6482	TGFb	43	-3.437	-3.714	0.00058	0.025
AZD6482	TGFb	43	-3.437	-3.714	0.00058	0.025
ZSTK474	TGFb	46	3.832	3.695	0.0006	0.015
GSK2126458	TNFa	26	-4.175	-3.945	0.00061	0.097
Cytarabine	TGFb	29	-7.218	-3.850	0.00069	0.097
AZD6482	p53	45	-3.470	-3.648	0.00071	0.025
AZD6482	p53	45	-3.470	-3.648	0.00071	0.025
5-Fluorouracil	EGFR	26	4.859	3.833	0.0008	0.098
ABT-263	MAPK	47	3.830	3.577	0.00083	0.018
Axitinib	Hypoxia	50	-2.714	-3.551	0.00087	0.018
BMS-708163	p53	50	-1.193	-3.536	0.00091	0.018

BMS-708163	p53	50	-1.193	-3.536	0.00091	0.018
Trametinib	JAK-STAT	47	6.506	3.487	0.0011	0.035
Trametinib	MAPK	47	5.721	3.380	0.0015	0.042
CI-1040	MAPK	44	4.189	3.368	0.0016	0.042
PD-0325901	p53	40	4.720	3.356	0.0018	0.044
ABT-263	TNFa	48	5.348	3.280	0.002	0.036
MP470	EGFR	44	4.063	3.242	0.0023	0.037
ABT-263	NFkB	48	6.189	3.217	0.0024	0.039
PD-0325901	VEGF	40	5.634	3.255	0.0024	0.052
MP470	p53	45	3.702	3.203	0.0026	0.037
CI-1040	p53	44	3.730	3.204	0.0026	0.052
Trametinib	TNFa	45	3.722	3.172	0.0027	0.052
ZSTK474	EGFR	45	3.372	3.183	0.0027	0.037
OSI-027	JAK-STAT	48	4.871	3.145	0.0028	0.043
AT-7519	MAPK	49	3.902	3.124	0.003	0.043
PD-0325901	TNFa	38	3.359	3.177	0.003	0.054
GSK2126458	EGFR	44	3.789	3.150	0.003	0.037
OSI-027	NFkB	51	5.496	3.075	0.0035	0.044
Afatinib	VEGF	43	4.527	3.086	0.0036	0.04
OSI-027	p53	50	-4.409	-3.040	0.0038	0.044
Axitinib	p53	50	-2.376	-3.041	0.0038	0.044
PD-0325901	EGFR	40	6.663	3.084	0.0039	0.056
Vorinostat	JAK-STAT	47	3.580	3.017	0.0041	0.045
RDEA119	MAPK	40	5.386	3.054	0.0042	0.056
CI-1040	TNFa	42	2.763	3.030	0.0042	0.056
Dasatinib	TNFa	16	-5.308	-3.391	0.0044	0.056
CI-1040	VEGF	44	4.076	3.004	0.0045	0.056
Bleomycin	p53	47	-5.864	-2.992	0.0045	0.056
Trametinib	p53	47	4.582	2.991	0.0045	0.056
MP470	TGFb	45	3.889	2.963	0.0049	0.045
Cytarabine	NFkB	44	-6.451	-2.972	0.0049	0.059
MP470	PI3K	44	3.627	2.961	0.005	0.045
PD-0325901	PI3K	40	6.271	2.949	0.0055	0.063
ABT-263	Hypoxia	48	3.916	2.855	0.0064	0.067
AT-7519	JAK-STAT	47	4.288	2.833	0.0067	0.067
RDEA119	TNFa	38	3.130	2.871	0.0067	0.071
Tipifarnib	p53	47	-3.571	-2.843	0.0067	0.071
ABT-263	TGFb	49	4.621	2.790	0.0076	0.069
Methotrexate	p53	50	-3.393	-2.784	0.0077	0.069
CEP-701	TNFa	42	2.140	2.801	0.0077	0.078
CH5424802	NFkB	51	1.842	2.767	0.008	0.069
ZSTK474	PI3K	45	2.891	2.772	0.0081	0.067
CI-1040	PI3K	44	4.623	2.761	0.0085	0.084
TG101348	p53	50	-2.278	-2.733	0.0088	0.072
AC220	Trail	49	-1.504	-2.710	0.0094	0.089
CEP-701	JAK-STAT	44	3.982	2.699	0.01	0.092
Methotrexate	JAK-STAT	47	4.877	2.635	0.011	0.086
AT-7519	VEGF	48	3.489	2.643	0.011	0.086
Dasatinib	VEGF	17	-11.610	-2.929	0.011	0.098
Masitinib	p53	49	-2.213	-2.632	0.011	0.099
Afatinib	TGFb	43	3.549	2.634	0.012	0.084
GSK2126458	PI3K	44	3.140	2.639	0.012	0.084
Afatinib (rescreen)	TGFb	45	3.731	2.572	0.014	0.09

Table B23: PARADIGM vs. drugs (tissue-specific)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
AC220	Trail	46	-834.661	-6.272	1.5e-07	1.3e-05
AC220	EGFR	20	-20.564	-7.168	8.2e-07	0.00089
AC220	EGFR	41	-29.924	-5.604	1.4e-06	6.2e-05
XL-184	TGFb	22	-0.903	-5.599	2.1e-05	0.011
Sorafenib	EGFR	19	-22.004	-4.910	0.00011	0.03
XL-880	TGFb	22	-0.778	-4.866	0.00011	0.03
Paclitaxel	TNFa	26	14.815	4.476	0.00017	0.15
Vinorelbine	VEGF	2	1158.739	4.493	0.00028	0.061
Nutlin-3a	Hypoxia	12	-0.720	-4.212	0.00039	0.061
XL-184	EGFR	20	-17.356	-4.288	0.0004	0.061
YM155	EGFR	35	58.933	3.794	0.00044	0.079
Sunitinib	EGFR	18	-18.552	-4.179	0.00063	0.078
PXD101, Belinostat	Trail	22	5.076	4.072	0.00065	0.078
Afatinib	MAPK	44	-1.037	-3.689	0.00066	0.02
PXD101, Belinostat	EGFR	37	41.576	3.512	0.00099	0.089

Table B24: Perturbation-response genes vs. drugs (tissue-specific)

Drug	Pathway	Size	Effect	Wald stat.	P-value	FDR
CH5424802	Hypoxia	27	-1.319	-5.416	1.5e-05	0.014
Bexarotene	PI3K	21	-1.303	-5.550	2.9e-05	0.034
ABT-263	JAK-STAT	49	2.237	4.427	5.8e-05	0.012
MP470	Trail	46	1.813	4.112	0.00017	0.017
AT-7519	p53	51	1.884	4.038	0.00019	0.019
BMS-708163	TGFb	26	-1.587	-4.344	0.00024	0.12
BMS-708163	TGFb	26	-1.587	-4.344	0.00024	0.12
OSI-027	Trail	22	-0.805	-4.491	0.00025	0.15
Cytarabine	VEGF	29	1.270	4.211	0.00027	0.12
Nilotinib	TNFa	10	3.552	6.593	0.00031	0.12
ABT-263	MAPK	49	1.219	3.391	0.0014	0.075
Axitinib	Hypoxia	51	0.549	3.339	0.0016	0.075
Vorinostat	JAK-STAT	51	1.000	3.253	0.0021	0.075
ABT-263	NFkB	49	1.972	3.234	0.0023	0.075
ABT-263	TNFa	49	1.577	3.169	0.0027	0.077

B.4 *Pathway scores and survival*

Pan-cancer

Table B25: Gene Ontology vs. patient survival (pan-cancer)

Pathway	Size	Effect	Wald stat.	P-value	FDR
Hypoxia	8922	0.917	5.606	2.1e-08	2.3e-07
p53	8922	0.308	3.302	0.00096	0.0053
NFkB	8922	0.225	3.116	0.0018	0.0067
TGFb	8922	0.394	2.943	0.0032	0.0089
MAPK	8922	0.526	2.747	0.006	0.013
TNFa	8922	0.227	2.526	0.012	0.021
Trail	8922	0.115	2.389	0.017	0.027
PI3K	8922	-0.128	-1.336	0.18	0.25
EGFR	8922	0.008	0.148	0.88	0.98
VEGF	8922	-0.003	-0.061	0.95	0.98
JAK-STAT	8922	0.002	0.019	0.98	0.98

Table B26: Reactome vs. patient survival (pan-cancer)

Pathway	Size	Effect	Wald stat.	P-value	FDR
TGFb	8922	0.486	4.112	3.9e-05	0.00043
VEGF	8922	0.204	1.731	0.084	0.28
PI3K	8922	-0.263	-1.586	0.11	0.28
p53	8922	0.144	1.536	0.12	0.28
Trail	8922	0.083	1.504	0.13	0.28
MAPK	8922	0.152	1.426	0.15	0.28
JAK-STAT	8922	0.097	0.973	0.33	0.48
Hypoxia	8922	0.096	0.938	0.35	0.48
EGFR	8922	0.068	0.447	0.66	0.8
NFkB	8922	0.020	0.218	0.83	0.91
TNFa	8922	-0.003	-0.032	0.97	0.97

Table B27: SPIA vs. patient survival (pan-cancer)

Pathway	Size	Effect	Wald stat.	P-value	FDR
EGFR	6501	0.004	3.997	6.4e-05	0.00051
JAK-STAT	6502	0.016	2.011	0.044	0.18
MAPK	6502	0.002	1.473	0.14	0.38
VEGF	6502	-0.003	-1.230	0.22	0.44
NFkB	6501	-0.001	-0.960	0.34	0.49
TGFb	6501	-0.002	-0.803	0.42	0.49
Trail	6502	0.001	0.790	0.43	0.49
PI3K	6501	-0.000	-0.050	0.96	0.96

Table B28: Pathifier vs. patient survival (pan-cancer)

Pathway	Size	Effect	Wald stat.	P-value	FDR
TNFa	6463	0.470	2.825	0.0047	0.021
EGFR	6494	0.517	2.815	0.0049	0.021
PI3K	6484	0.500	2.768	0.0056	0.021
MAPK	6490	0.407	2.286	0.022	0.061
Trail	6462	0.282	1.805	0.071	0.16
p53	6461	0.281	1.658	0.097	0.18
TGFb	6476	-0.294	-1.538	0.12	0.19
JAK-STAT	6483	-0.121	-0.754	0.45	0.59
Hypoxia	6475	0.122	0.706	0.48	0.59
VEGF	6485	-0.022	-0.121	0.9	0.94
NFkB	6484	-0.012	-0.074	0.94	0.94

Table B29: PARADIGM vs. patient survival (pan-cancer)

Pathway	Size	Effect	Wald stat.	P-value	FDR
EGFR	7494	0.697	3.150	0.0016	0.016
PI3K	5835	0.032	2.365	0.018	0.084
Hypoxia	4463	0.039	2.231	0.026	0.084
TGFb	8784	0.036	2.126	0.034	0.084
p53	7968	0.064	2.032	0.042	0.084
TNFa	8146	0.479	1.837	0.066	0.11
MAPK	8784	0.021	0.849	0.4	0.57
JAK-STAT	8784	0.013	0.682	0.5	0.62
VEGF	1615	-0.176	-0.416	0.68	0.75
Trail	8784	-0.093	-0.163	0.87	0.87

Table B30: Perturbation-response genes vs. patient survival (pan-cancer)

Pathway	Size	Effect	Wald stat.	P-value	FDR
EGFR	8922	0.296	8.456	0	0
PI3K	8922	0.415	8.302	1.1e-16	6.1e-16
MAPK	8922	0.307	8.085	6.7e-16	2.4e-15
Hypoxia	8922	0.250	6.091	1.1e-09	3.1e-09
TGFb	8922	0.182	4.716	2.4e-06	5.3e-06
TNFa	8922	0.157	4.330	1.5e-05	2.7e-05
Trail	8922	-0.069	-2.230	0.026	0.04
NFkB	8922	0.051	1.480	0.14	0.19
p53	8922	-0.062	-1.387	0.17	0.2
VEGF	8922	0.036	1.142	0.25	0.28
JAK-STAT	8922	0.014	0.480	0.63	0.63

Tissue specific

Table B31: Gene Ontology vs. patient survival (tissue-specific)

tissue	Pathway	Size	Effect	Wald stat.	P-value	FDR
LGG	JAK-STAT	495	2.475	6.160	7.3e-10	8e-09
KIRC	VEGF	519	-0.971	-4.944	7.7e-07	8.4e-06
LGG	Trail	495	1.169	4.903	9.4e-07	5.2e-06
KIRC	EGFR	519	-0.822	-4.731	2.2e-06	1.2e-05
CESC	TGFb	277	4.003	4.651	3.3e-06	3.6e-05
PAAD	p53	174	2.441	4.479	7.5e-06	8.2e-05
KIRC	TGFb	519	-1.586	-3.761	0.00017	0.00055
KIRC	PI3K	519	-1.398	-3.720	0.0002	0.00055
PAAD	Trail	174	0.957	3.723	0.0002	0.0011
HNSC	Trail	509	0.648	3.655	0.00026	0.0028
SARC	JAK-STAT	253	-1.782	-3.598	0.00032	0.0035
LGG	p53	495	1.720	3.490	0.00048	0.0018
KIRP	p53	267	3.085	3.442	0.00058	0.0064
CESC	Hypoxia	277	3.207	3.368	0.00076	0.0042
LGG	TNFa	495	1.528	3.229	0.0012	0.0027
LGG	Hypoxia	495	2.734	3.232	0.0012	0.0027
KIRC	TNFa	519	0.958	3.081	0.0021	0.0045
PAAD	TGFb	174	2.443	3.046	0.0023	0.0085

MESO	TGFb	80	2.197	2.962	0.0031	0.034
ACC	EGFR	79	-1.725	-2.918	0.0035	0.039
LUSC	TGFb	469	1.410	2.806	0.005	0.055
KIRC	Hypoxia	519	-1.418	-2.766	0.0057	0.01
KIRP	TGFb	267	2.632	2.710	0.0067	0.037
LGG	TGFb	495	1.363	2.560	0.01	0.019
PAAD	MAPK	174	2.720	2.549	0.011	0.03
HNSC	NFkB	509	0.595	2.490	0.013	0.07
PAAD	Hypoxia	174	2.113	2.471	0.013	0.03
KIRC	NFkB	519	0.571	2.378	0.017	0.027
MESO	JAK-STAT	80	-1.203	-2.357	0.018	0.07
MESO	Hypoxia	80	2.395	2.344	0.019	0.07
PAAD	NFkB	174	0.982	2.340	0.019	0.031
PAAD	TNFa	174	1.224	2.327	0.02	0.031
MESO	VEGF	80	0.861	2.199	0.028	0.077
PAAD	PI3K	174	0.957	1.943	0.052	0.072

Table B32: Reactome vs. patient survival (tissue-specific)

tissue	Pathway	Size	Effect	Wald stat.	P-value	FDR
LGG	TNFa	495	2.046	5.417	6.1e-08	6.7e-07
LGG	JAK-STAT	495	2.154	5.019	5.2e-07	2.9e-06
KIRC	VEGF	519	-1.975	-4.993	5.9e-07	6.5e-06
PAAD	TGFb	174	3.340	4.888	1e-06	1.1e-05
LGG	Hypoxia	495	-2.497	-4.584	4.6e-06	1.3e-05
LGG	Trail	495	1.111	4.580	4.7e-06	1.3e-05
KIRC	TGFb	519	-1.735	-4.398	1.1e-05	6e-05
SARC	NFkB	253	-1.978	-4.045	5.2e-05	0.00057
LGG	NFkB	495	1.695	4.030	5.6e-05	0.00012
KIRC	PI3K	519	-2.301	-3.825	0.00013	0.00048
KIRC	MAPK	519	-1.244	-3.638	0.00027	0.00066
KIRC	EGFR	519	-1.958	-3.615	0.0003	0.00066
LGG	p53	495	-1.862	-3.604	0.00031	0.00057
PAAD	Hypoxia	174	2.220	3.554	0.00038	0.0017
KIRC	Hypoxia	519	-1.312	-3.503	0.00046	0.00084
PAAD	Trail	174	0.995	3.498	0.00047	0.0017
ACC	TGFb	79	3.876	3.396	0.00068	0.0075
PAAD	JAK-STAT	174	1.971	3.259	0.0011	0.0031
SARC	JAK-STAT	253	-1.650	-3.263	0.0011	0.004
SARC	TNFa	253	-1.558	-3.277	0.0011	0.004
HNSC	p53	509	1.046	3.195	0.0014	0.015
PAAD	TNFa	174	1.365	2.911	0.0036	0.0079
LUAD	TGFb	462	1.353	2.881	0.004	0.044
PAAD	MAPK	174	1.556	2.824	0.0047	0.0087
SARC	Trail	253	-0.760	-2.810	0.0049	0.014
ACC	PI3K	79	-5.621	-2.817	0.0049	0.027
CESC	TGFb	277	1.755	2.709	0.0067	0.074
BRCA	Hypoxia	1018	1.283	2.639	0.0083	0.065
LGG	TGFb	495	1.317	2.629	0.0086	0.013
BRCA	p53	1018	1.086	2.516	0.012	0.065
PAAD	NFkB	174	1.340	2.472	0.013	0.019
PAAD	VEGF	174	1.509	2.459	0.014	0.019
LGG	VEGF	495	1.327	2.442	0.015	0.02
ACC	p53	79	-2.460	-2.236	0.025	0.093

Table B33: SPIA vs. patient survival (tissue-specific)

tissue	Pathway	Size	Effect	Wald stat.	P-value	FDR
KIRC	EGFR	519	-0.015	-4.699	2.6e-06	2.1e-05
KIRC	VEGF	519	-0.032	-4.228	2.4e-05	9.4e-05
KIRC	PI3K	519	-0.083	-3.894	9.9e-05	0.00026
PAAD	EGFR	174	0.019	3.630	0.00028	0.0023
CESC	EGFR	277	0.016	3.545	0.00039	0.0031
BRCA	JAK-STAT	1018	0.104	3.533	0.00041	0.0033
KIRC	TGFb	519	-0.018	-2.925	0.0034	0.0069
CESC	MAPK	277	0.014	2.898	0.0038	0.015
HNSC	EGFR	509	0.006	2.693	0.0071	0.057
PAAD	Trail	174	0.016	2.562	0.01	0.042
CESC	Trail	277	0.014	2.461	0.014	0.037
KIRC	MAPK	519	0.010	2.085	0.037	0.045

KIRC	NFkB	519	0.007	2.077	0.038	0.045
KIRC	Trail	519	0.010	2.061	0.039	0.045

Table B34: Pathifier vs. patient survival (tissue-specific)

tissue	Pathway	Size	Effect	Wald stat.	P-value	FDR
KIRC	NFkB	517	4.484	6.528	6.7e-11	7.3e-10
KIRC	PI3K	514	4.508	6.236	4.5e-10	2.5e-09
KIRC	EGFR	513	3.767	5.244	1.6e-07	5.8e-07
KIRC	TGFb	518	4.204	4.997	5.8e-07	1.6e-06
KIRC	p53	518	4.340	4.605	4.1e-06	9.1e-06
KIRC	JAK-STAT	510	2.602	4.300	1.7e-05	3.1e-05
KIRC	MAPK	517	3.501	4.189	2.8e-05	4.4e-05
PAAD	Trail	172	3.122	3.436	0.00059	0.0065
CESC	TNFa	274	2.991	3.364	0.00077	0.0084
HNSC	TNFa	501	1.261	3.051	0.0023	0.025
PAAD	TGFb	174	-4.152	-3.014	0.0026	0.014
KIRC	TNFa	517	2.185	2.998	0.0027	0.0037
LUSC	NFkB	469	-1.360	-2.871	0.0041	0.045
PAAD	JAK-STAT	174	-2.928	-2.827	0.0047	0.014
PAAD	NFkB	171	-2.669	-2.800	0.0051	0.014
CESC	TGFb	277	-2.288	-2.675	0.0075	0.041
KIRC	VEGF	514	-1.836	-2.522	0.012	0.014
HNSC	Trail	507	1.084	2.450	0.014	0.079
COAD	TNFa	341	1.692	2.214	0.027	0.08
COAD	NFkB	341	1.391	2.127	0.033	0.08
PAAD	MAPK	170	1.408	2.119	0.034	0.075
COAD	TGFb	341	1.435	1.973	0.048	0.08
COAD	MAPK	341	1.480	1.953	0.051	0.08
COAD	p53	341	1.271	1.946	0.052	0.08
COAD	Hypoxia	341	1.180	1.934	0.053	0.08
COAD	JAK-STAT	341	1.363	1.906	0.057	0.08
COAD	EGFR	341	1.293	1.874	0.061	0.08
COAD	PI3K	341	1.340	1.843	0.065	0.08
COAD	Trail	341	1.327	1.721	0.085	0.086
COAD	VEGF	341	1.306	1.718	0.086	0.086

Table B35: PARADIGM vs. patient survival (tissue-specific)

tissue	Pathway	Size	Effect	Wald stat.	P-value	FDR
LGG	JAK-STAT	495	0.344	4.124	3.7e-05	0.00024
ACC	TNFa	73	8.767	4.125	3.7e-05	0.00037
LGG	EGFR	394	2.558	4.068	4.7e-05	0.00024
PAAD	PI3K	117	0.264	3.521	0.00043	0.0043
BRCA	Trail	1018	7.516	3.371	0.00075	0.0075
STAD	TGFb	366	0.259	3.201	0.0014	0.014
LGG	MAPK	495	0.350	3.165	0.0016	0.0052
KIRC	TGFb	381	-0.230	-2.978	0.0029	0.029
UCEC	Hypoxia	305	0.282	2.874	0.004	0.04
BRCA	PI3K	663	0.161	2.740	0.0061	0.031
LUSC	JAK-STAT	469	0.168	2.684	0.0073	0.073
KIRC	PI3K	257	-0.152	-2.670	0.0076	0.038
KIRP	EGFR	214	6.056	2.622	0.0087	0.079
LUAD	PI3K	308	0.138	2.600	0.0093	0.093
HNSC	Hypoxia	258	-0.186	-2.594	0.0095	0.095
PAAD	p53	160	0.477	2.524	0.012	0.058
LGG	TNFa	452	3.080	2.394	0.017	0.042
KIRP	MAPK	267	0.372	2.344	0.019	0.079
KIRP	TGFb	267	0.299	2.263	0.024	0.079

Table B36: Perturbation-response genes vs. patient survival (tissue-specific)

tissue	Pathway	Size	Effect	Wald stat.	P-value	FDR
KIRC	TNFa	519	0.828	5.182	2.2e-07	2.4e-06
ACC	MAPK	79	1.375	4.833	1.3e-06	1.5e-05
LGG	JAK-STAT	495	0.533	4.808	1.5e-06	1.4e-05
LGG	TNFa	495	0.813	4.700	2.6e-06	1.4e-05
KIRC	NFkB	519	0.654	4.422	9.8e-06	5.4e-05

PAAD	MAPK	174	1.153	4.393	1.1e-05	0.00012
KIRC	EGFR	519	0.610	4.252	2.1e-05	7.8e-05
ACC	EGFR	79	0.969	4.192	2.8e-05	0.00015
KIRP	PI3K	267	1.826	4.034	5.5e-05	0.0006
ACC	PI3K	79	2.039	4.026	5.7e-05	0.00021
LIHC	Hypoxia	333	0.703	4.006	6.2e-05	0.00068
CESC	Hypoxia	277	0.939	3.998	6.4e-05	0.0007
LGG	MAPK	495	0.672	3.932	8.4e-05	0.00031
ACC	p53	79	-1.395	-3.865	0.00011	0.00031
LUAD	PI3K	462	0.854	3.774	0.00016	0.0013
CESC	MAPK	277	0.743	3.753	0.00018	0.00096
LGG	NFkB	495	0.613	3.737	0.00019	0.00051
MESO	PI3K	80	1.112	3.714	0.0002	0.0022
LUAD	EGFR	462	0.530	3.680	0.00023	0.0013
LGG	EGFR	495	0.671	3.677	0.00024	0.00052
BRCA	PI3K	1018	0.693	3.582	0.00034	0.0037
CESC	EGFR	277	0.600	3.383	0.00072	0.0026
UCEC	VEGF	531	-0.751	-3.375	0.00074	0.0081
SARC	JAK-STAT	253	-0.535	-3.327	0.00088	0.0058
PAAD	EGFR	174	0.675	3.319	0.0009	0.005
LUAD	MAPK	462	0.519	3.297	0.00098	0.0036
SARC	Trail	253	-0.541	-3.273	0.0011	0.0058
CESC	TNFa	277	0.755	3.237	0.0012	0.0033
KICH	Hypoxia	64	2.730	3.181	0.0015	0.016
PAAD	JAK-STAT	174	0.572	3.157	0.0016	0.0058
KIRC	PI3K	519	0.648	3.041	0.0024	0.0065
LUAD	Hypoxia	462	0.502	2.961	0.0031	0.0084
PRAD	Trail	472	-2.169	-2.906	0.0037	0.04
CESC	TGFb	277	0.700	2.898	0.0038	0.0083
SARC	NFkB	253	-0.492	-2.883	0.0039	0.012
KIRP	EGFR	267	0.786	2.860	0.0042	0.017
SARC	MAPK	253	0.479	2.850	0.0044	0.012
KIRP	MAPK	267	0.683	2.823	0.0048	0.017
PAAD	Hypoxia	174	0.507	2.788	0.0053	0.015
LGG	PI3K	495	0.741	2.714	0.0066	0.012
HNSC	MAPK	509	0.358	2.716	0.0066	0.073
MESO	TGFb	80	0.549	2.705	0.0068	0.038
UVM	PI3K	77	1.681	2.700	0.0069	0.076
CESC	PI3K	277	0.843	2.667	0.0076	0.014
KIRP	TGFb	267	0.913	2.627	0.0086	0.024
LIHC	MAPK	333	0.511	2.616	0.0089	0.044
LGG	TGFb	495	0.856	2.592	0.0095	0.015
LIHC	PI3K	333	0.541	2.514	0.012	0.044
PAAD	TNFa	174	0.473	2.494	0.013	0.025
KICH	PI3K	64	4.270	2.454	0.014	0.067
PAAD	PI3K	174	0.945	2.464	0.014	0.025
ACC	TGFb	79	0.942	2.437	0.015	0.03
MESO	EGFR	80	0.382	2.407	0.016	0.059
ACC	Trail	79	-0.940	-2.395	0.017	0.03
KICH	p53	64	-1.991	-2.363	0.018	0.067
LGG	Trail	495	0.365	2.274	0.023	0.032
SARC	Hypoxia	253	0.401	2.223	0.026	0.049
CESC	Trail	277	-0.407	-2.232	0.026	0.04
SARC	TNFa	253	-0.384	-2.213	0.027	0.049
MESO	JAK-STAT	80	-0.348	-2.195	0.028	0.077
KIRC	JAK-STAT	519	0.419	2.174	0.03	0.065
PAAD	p53	174	0.838	2.074	0.038	0.06
KIRP	Hypoxia	267	0.468	1.969	0.049	0.093
KIRP	NFkB	267	0.641	1.956	0.051	0.093
PAAD	NFkB	174	0.383	1.910	0.056	0.077
PAAD	TGFb	174	0.412	1.822	0.068	0.084

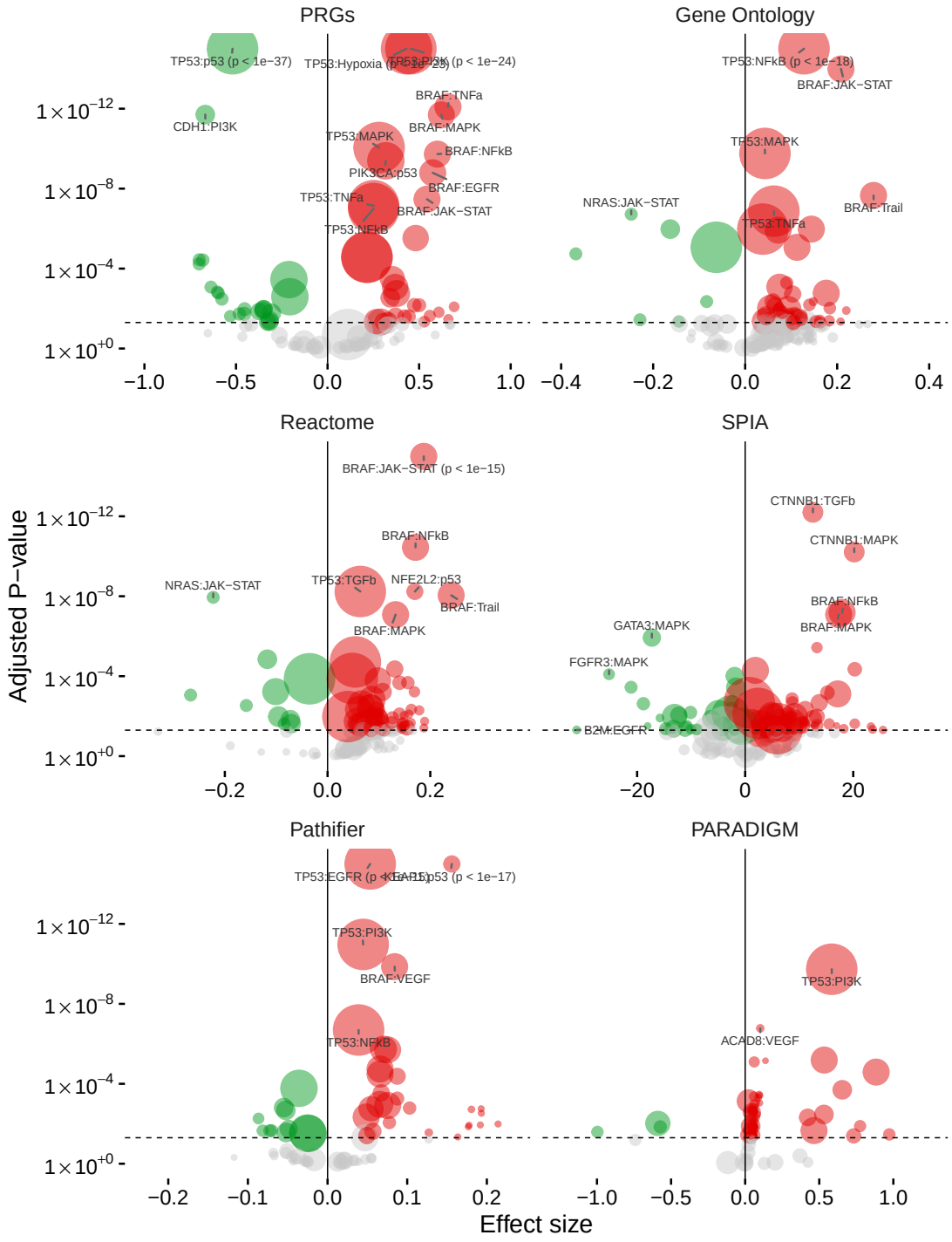

Figure B1: Volcano plot for pan-cancer associations between pathway scores and mutated driver genes. Effect size is standard deviations of pathway scores. P-values FDR-corrected. Associations corrected for can- cer type. PRGs provide stronger associations and are more in line with literature knowledge of signaling pathways.

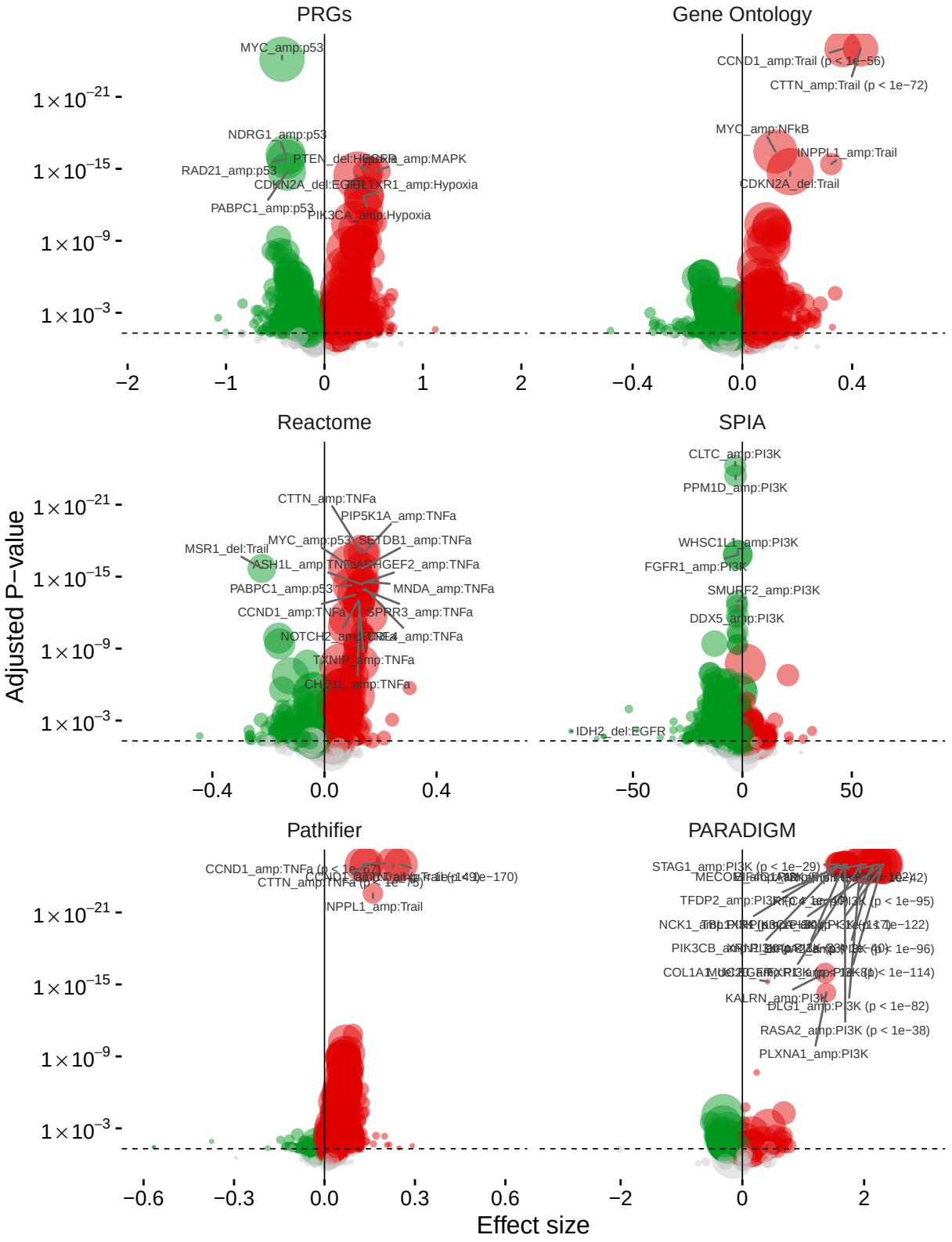

Figure B2: Volcano plots for pan-cancer associations between pathway scores and GISTIC-filtered gene ampli- fications and deletion. Effect size is standard deviations of pathway scores. P-values FDR-corrected. Cancer type regressed out. Associations by PRGs are more in line with established literature knowledge.

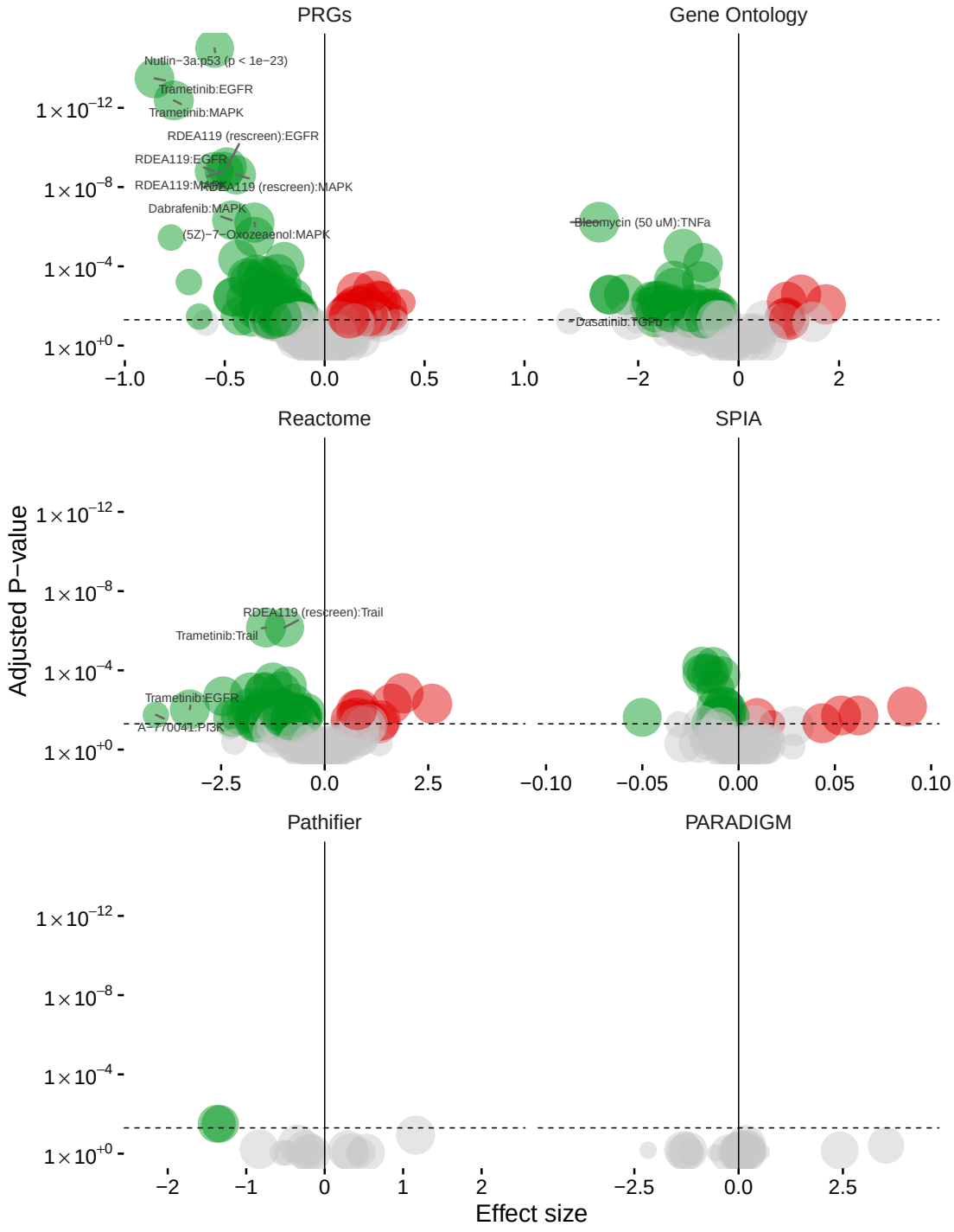

Figure B3: Pan-cancer volcano plots for associations between pathway scores and drug response (IC50). Effect size arbitrary units, p-values FDR-corrected. Pathway-response genes are the only method to recover highly significant oncogene addiction associations, the rest of methods show no obvious connection between the drug target and pathway.

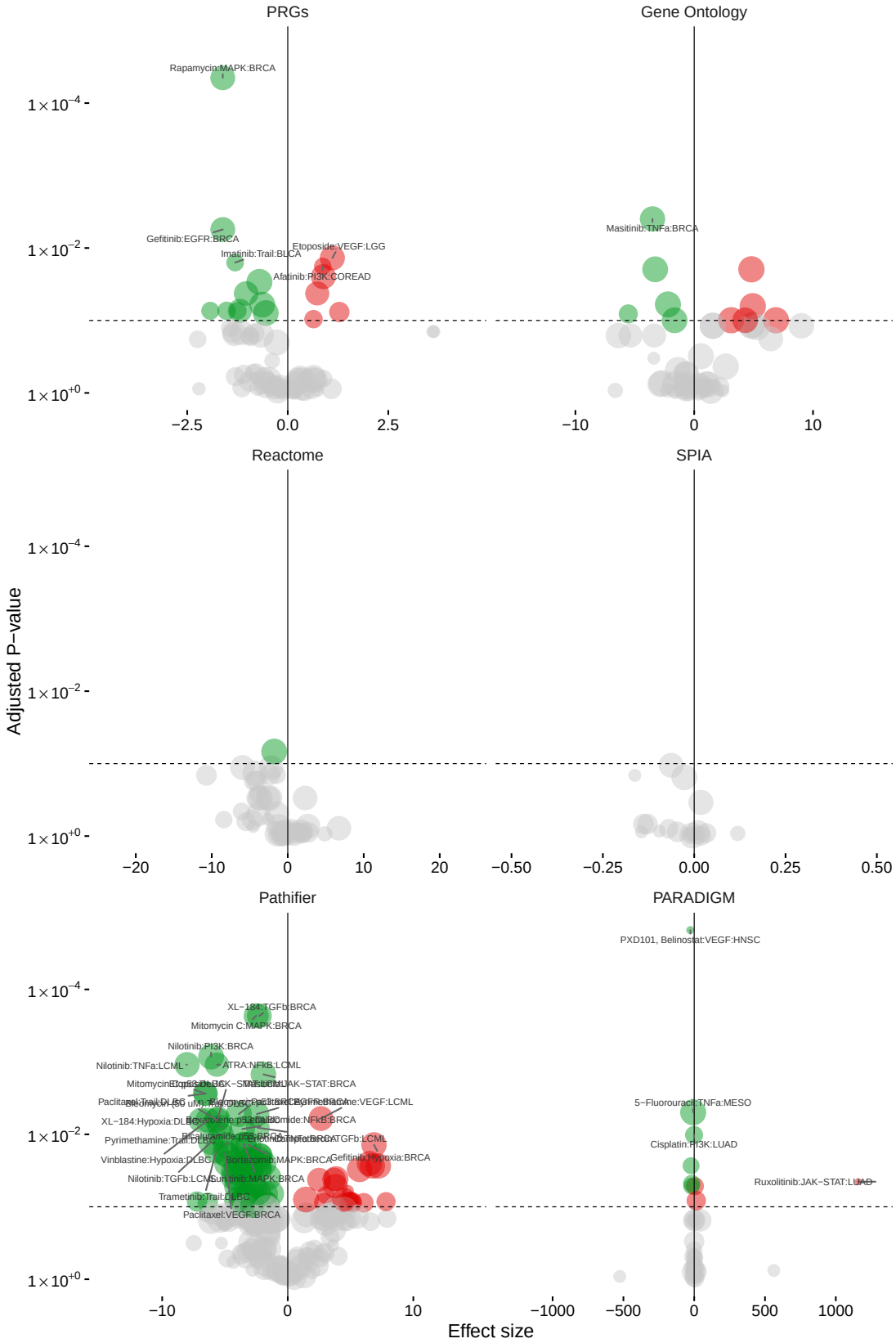

Figure B4: Tissue-specific volcano plots for associations between pathway scores and drug response (IC50). Effect size arbitrary units, p-values FDR-corrected. Pathway-response genes are the only method to recover highly significant oncogene addiction associations, the rest of methods show no obvious connection between the drug target and pathway.

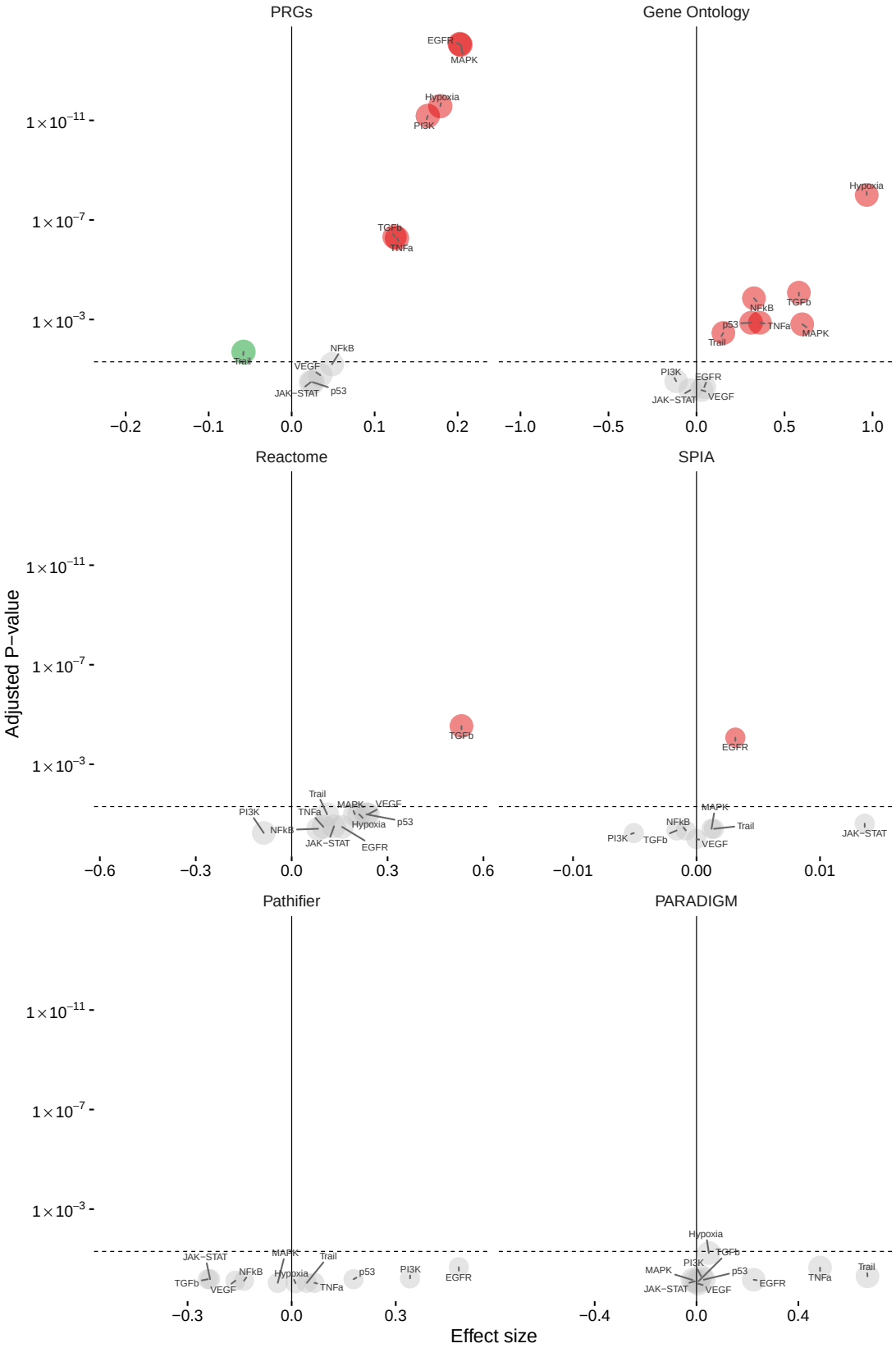

Figure B5: Volcano plots for pan-cancer survival associations. Effect size arbitrary units, p-values FDR- corrected. Pathway-response genes provide stronger associations and are the only method to separate associations into classical oncogenic and tumor suppressor pathways, calling into question the meaning of associations obtained by other methods.

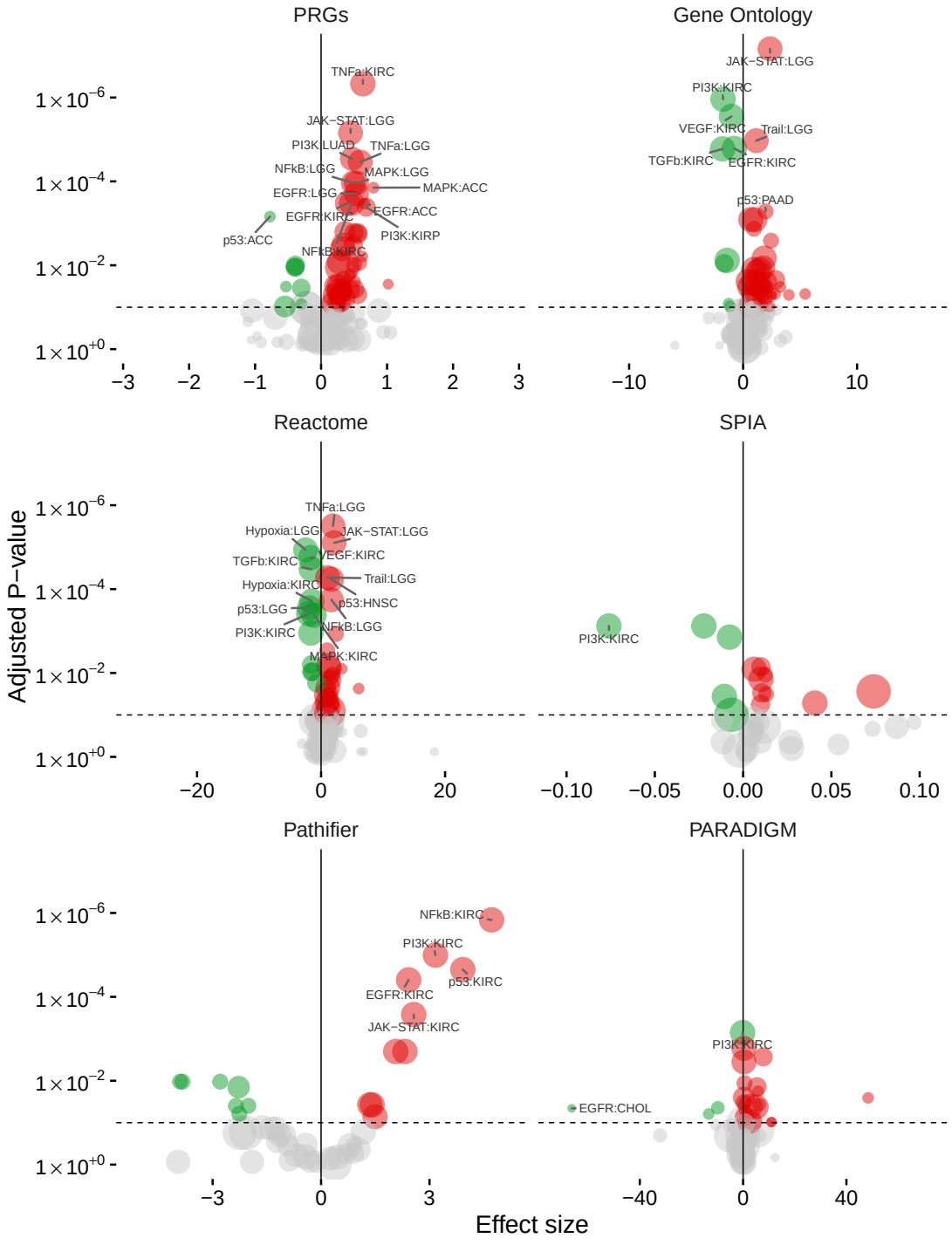

Figure B6: Volcano plots for tissue-specific survival associations. Effect size arbitrary units, p-values FDR- corrected. All methods show strongest associations with KIRC and LGG. Pathway-response genes only method to separate associations into classical oncogenic and tumor suppressor pathways, calling into question the meaning of associations obtained by other methods.

C DRUG SENSITISATION (CHAPTER 5)

C.1 *MANTRA*

Pan-cancer

Table C1: MANTRA (pan-cancer)

Drug response	Pathway	Size	Effect	Wald stat.	P-value	FDR
NPK76-II-72-1	cyclic_adenosine_monophosphate	733	-0.448	-6.698	4.3e-11	9.3e-06
NPK76-II-72-1	bromocriptine	733	0.358	6.664	5.4e-11	9.3e-06
Bleomycin (50 uM)	diphemanil_metilsulfate	746	-0.509	-6.537	1.2e-10	1.4e-05
Lenalidomide	fenoterol	677	0.190	6.453	2.1e-10	1.9e-05
NPK76-II-72-1	1,4-chrysenequinone	733	0.395	6.375	3.3e-10	2.3e-05
Lenalidomide	methylergometrine	677	0.179	6.299	5.5e-10	3.2e-05
NPK76-II-72-1	chlorogenic_acid	733	-0.409	-6.252	7e-10	3.2e-05
CCT018159	naproxen	715	0.269	6.240	7.7e-10	3.2e-05
GSK-650394	mepacrine	680	0.435	6.222	8.8e-10	3.2e-05
Lenalidomide	tetryzoline	677	0.192	6.194	1e-09	3.2e-05
GSK429286A	1,4-chrysenequinone	734	0.274	6.169	1.2e-09	3.2e-05
QS11	STOCK1N-35874	689	-0.339	-6.157	1.3e-09	3.2e-05
Docetaxel	clidinium_bromide	676	0.440	6.153	1.3e-09	3.2e-05
GSK429286A	N6-methyladenosine	734	0.302	6.114	1.6e-09	3.2e-05
NPK76-II-72-1	daunorubicin	733	0.374	6.119	1.6e-09	3.2e-05
NPK76-II-72-1	albendazole	733	0.350	6.107	1.7e-09	3.2e-05
NPK76-II-72-1	dilazep	733	0.334	6.103	1.7e-09	3.2e-05
NPK76-II-72-1	pirlindole	733	0.327	6.110	1.7e-09	3.2e-05
Lenalidomide	quinostatin	677	0.174	6.108	1.7e-09	3.2e-05
GSK-650394	isoetarine	680	0.406	6.099	1.8e-09	3.2e-05
NPK76-II-72-1	pimethixene	733	0.423	6.068	2.1e-09	3.4e-05
EX-527	1,4-chrysenequinone	731	0.165	6.063	2.2e-09	3.4e-05
FK866	carbinoxamine	718	0.637	6.039	2.5e-09	3.6e-05
Lenalidomide	sulfasalazine	677	0.183	6.045	2.5e-09	3.6e-05
QS11	isoetarine	689	0.341	6.040	2.6e-09	3.6e-05
Lenalidomide	bromocriptine	677	0.175	6.023	2.9e-09	3.8e-05
T0901317	N6-methyladenosine	728	0.244	5.992	3.3e-09	4e-05
EX-527	ozagrel	731	-0.244	-5.995	3.3e-09	4e-05
Lenalidomide	wortmannin	677	0.169	5.998	3.3e-09	4e-05
NPK76-II-72-1	fulvestrant	733	0.363	5.962	3.9e-09	4.5e-05
GSK1070916	1,4-chrysenequinone	713	0.440	5.963	4e-09	4.5e-05
EHT 1864	ritodrine	737	0.204	5.942	4.4e-09	4.8e-05
ZM-447439	clidinium_bromide	635	0.382	5.949	4.6e-09	4.8e-05
Vismodegib	cyclic_adenosine_monophosphate	676	-0.241	-5.907	5.6e-09	5.6e-05
Trametinib	PF-00875133-00	713	-0.563	-5.896	5.9e-09	5.6e-05
EHT 1864	bacampicillin	737	-0.211	-5.890	6e-09	5.6e-05
EHT 1864	neostigmine_bromide	737	-0.208	-5.889	6e-09	5.6e-05
Lenalidomide	perhexiline	677	0.170	5.888	6.3e-09	5.7e-05
Bleomycin (50 uM)	STOCK1N-35696	746	-0.422	-5.872	6.6e-09	5.9e-05
Lenalidomide	pentamidine	677	0.197	5.872	6.9e-09	6e-05
NPK76-II-72-1	(-)-isoprenaline	733	0.339	5.848	7.6e-09	6.1e-05
Lenalidomide	fulvestrant	677	0.191	5.852	7.7e-09	6.1e-05
NPK76-II-72-1	tranylcypromine	733	0.332	5.841	7.9e-09	6.1e-05
Lenalidomide	etacrynic_acid	677	0.181	5.846	8e-09	6.1e-05
Gemcitabine	1,4-chrysenequinone	685	0.708	5.842	8.1e-09	6.1e-05
Cytarabine	cyclic_adenosine_monophosphate	673	-0.456	-5.844	8.1e-09	6.1e-05
Lenalidomide	latamoxef	677	0.165	5.830	8.8e-09	6.5e-05
Bleomycin (50 uM)	domperidone	746	-0.432	-5.812	9.3e-09	6.7e-05
Afatinib (rescreen)	hydroxyachillin	727	0.565	5.809	9.5e-09	6.7e-05
EX-527	metoprolol	731	0.196	5.807	9.6e-09	6.7e-05

Tissue specific

Table C2: MANTRA (cancer-specific)

Drug response	Pathway	Size	Effect	Wald stat.	P-value	FDR
BAY 61-3606	cisapride	39	1.517	8.057	1.2e-09	0.0004
Lenalidomide	pheniramine	51	1.417	7.150	3.9e-09	0.00067
BAY 61-3606	halofantrine	39	0.986	7.102	2.1e-08	0.0013

Shikonin	cisapride	39	1.178	7.084	2.2e-08	0.0013
PD-173074	pentolonium	51	1.105	6.653	2.3e-08	0.0013
Lenalidomide	harpagoside	51	1.458	6.624	2.5e-08	0.0013
Shikonin	thioridazine	39	1.297	6.977	3e-08	0.0013
Lenalidomide	isoxsuprine	51	1.576	6.542	3.4e-08	0.0013
PD-173074	scriptaid	51	1.287	6.532	3.5e-08	0.0013
Shikonin	reserpine	39	0.952	6.898	3.9e-08	0.0013
Shikonin	lobeline	39	0.817	6.773	5.7e-08	0.0016
Shikonin	methoxsalen	39	1.065	6.773	5.7e-08	0.0016
Lenalidomide	hyoscyamine	51	1.383	6.376	6.2e-08	0.0016
PD-173074	delsoline	51	1.382	6.331	7.2e-08	0.0017
PD-173074	nalidixic_acid	51	1.294	6.333	7.2e-08	0.0017
BAY 61-3606	amantadine	39	1.025	6.629	8.9e-08	0.0019
Shikonin	halofantrine	39	0.778	6.579	1e-07	0.0021
BAY 61-3606	5152487	39	-0.947	-6.547	1.1e-07	0.0021
Lenalidomide	dopamine	51	1.574	6.217	1.1e-07	0.0021
Shikonin	tolazoline	39	1.027	6.530	1.2e-07	0.022
Shikonin	vorinostat	39	1.048	6.473	1.4e-07	0.022
Shikonin	dopamine	39	1.157	6.487	1.4e-07	0.0024
Shikonin	5152487	39	-0.767	-6.424	1.7e-07	0.0027
Shikonin	amantadine	39	0.824	6.407	1.8e-07	0.0027
Lenalidomide	nalidixic_acid	51	1.479	6.076	1.8e-07	0.0027
Lenalidomide	tiletamine	51	1.326	6.057	1.9e-07	0.022
PD-173074	betahistine	51	1.169	6.064	1.9e-07	0.0027
PD-173074	epirizole	51	1.112	6.000	2.3e-07	0.0028
BAY 61-3606	methoxsalen	39	1.261	6.308	2.4e-07	0.0028
PD-173074	sulfamerazine	51	1.171	5.997	2.4e-07	0.0028
Lenalidomide	butirosin	51	1.554	5.980	2.5e-07	0.0028
Lenalidomide	canavanine	51	1.263	5.980	2.5e-07	0.0028
Tamoxifen	canavanine	51	0.592	5.985	2.5e-07	0.0028
PD-173074	dihydroergotamine	51	1.178	5.975	2.6e-07	0.0028
Shikonin	idazoxan	39	0.949	6.282	2.6e-07	0.0028
Lenalidomide	probucol	51	1.102	5.950	2.8e-07	0.0029
PD-173074	indometacin	51	1.192	5.923	3.1e-07	0.0031
PD-173074	budesonide	51	1.080	5.913	3.2e-07	0.0032
Shikonin	withaferin_A	39	0.980	6.210	3.3e-07	0.028
PHA-665752	sulfamethoxazole	25	1.098	7.053	3.5e-07	0.0033
AG-014699	nalidixic_acid	48	0.844	5.915	3.9e-07	0.0035
Lenalidomide	sulindac	51	1.309	5.859	3.9e-07	0.0035
PD-173074	dicoumarol	51	1.252	5.839	4.1e-07	0.0037
Lenalidomide	etidronic_acid	51	1.178	5.819	4.4e-07	0.0037
AC220	probucol	50	1.160	5.842	4.4e-07	0.0037
Lenalidomide	dicoumarol	51	1.577	5.812	4.5e-07	0.0037
Lenalidomide	delsoline	51	1.613	5.807	4.6e-07	0.0037
Shikonin	isoxsuprine	39	1.076	6.086	4.8e-07	0.0037
PD-173074	lactobionic_acid	51	1.115	5.800	4.8e-07	0.0037
Lenalidomide	vinburnine	51	1.231	5.793	4.9e-07	0.034

C.2 *LINCS Connectivity Map*

Naive

Table C3: Drug sensitisation (naive pan-cancer)

Drug response	Sensitiser	Effect	Wald stat.	P-value	FDR
Afatinib (rescreen)	BMS-536924	-0.451	-6.561	1e-10	3.4e-06
Afatinib	PI-103	-0.419	-6.469	2e-10	3.4e-06
Afatinib	BMS-536924	-0.412	-6.349	4.1e-10	3.9e-06
Afatinib	NVP-BEZ235	-0.439	-6.332	4.5e-10	3.9e-06
Afatinib	Dasatinib	-0.373	-5.935	4.8e-09	3.1e-05
Afatinib	ZSTK474	-0.399	-5.915	5.4e-09	3.1e-05
Afatinib	GSK2126458	-0.361	-5.710	1.7e-08	6.8e-05
Afatinib (rescreen)	NVP-TAE684	-0.447	-5.703	1.7e-08	6.8e-05
Afatinib	NVP-TAE684	-0.418	-5.704	1.8e-08	6.8e-05
Afatinib	KIN001-102	-0.352	-5.681	2e-08	6.9e-05
Afatinib (rescreen)	AP-24534	-0.391	-5.635	2.5e-08	6.9e-05
Afatinib (rescreen)	PI-103	-0.391	-5.641	2.5e-08	6.9e-05
Gefitinib	NVP-BEZ235	-0.292	-5.624	2.8e-08	6.9e-05
Afatinib (rescreen)	NVP-BEZ235	-0.415	-5.616	2.8e-08	6.9e-05
Afatinib (rescreen)	Dasatinib	-0.374	-5.541	4.3e-08	9.8e-05
Afatinib (rescreen)	ZSTK474	-0.391	-5.495	5.5e-08	0.00012

Afatinib	AZD6482	-0.364	-5.452	7.1e-08	0.00014
Gefitinib	PI-103	-0.263	-5.410	8.9e-08	0.00017
Gefitinib	BMS-536924	-0.261	-5.382	1e-07	0.00017
Gefitinib	Dasatinib	-0.254	-5.386	1e-07	0.00017
Afatinib	AZD8055	-0.361	-5.381	1e-07	0.00017
Afatinib	AP-24534	-0.348	-5.323	1.4e-07	0.00022
Gefitinib	KIN001-102	-0.243	-5.291	1.7e-07	0.00024
Afatinib (rescreen)	GSK2126458	-0.356	-5.288	1.7e-07	0.00024
Afatinib	GDC0941	-0.332	-5.256	2e-07	0.00028
Gefitinib	ZSTK474	-0.265	-5.249	2.1e-07	0.00028
Afatinib (rescreen)	Embelin	-0.426	-5.231	2.2e-07	0.00029
Afatinib	PLX4720	-0.362	-5.129	3.9e-07	0.00046
Afatinib	PLX4720	-0.362	-5.129	3.9e-07	0.00046
Afatinib (rescreen)	FR-180204	-0.337	-5.108	4.2e-07	0.00048
Docetaxel	MS-275	0.313	5.088	4.8e-07	0.00053
Afatinib	OSI-027	-0.315	-5.027	6.5e-07	0.0007
Afatinib	BMS-754807	-0.301	-5.020	6.7e-07	0.0007
Gefitinib	GSK2126458	-0.233	-4.938	1e-06	0.001
Cisplatin	HG-6-64-1	0.225	4.880	1.3e-06	0.0013
Gefitinib	PLX4720	-0.256	-4.874	1.4e-06	0.0013
Gefitinib	PLX4720	-0.256	-4.874	1.4e-06	0.0013
Afatinib (rescreen)	AZD8055	-0.344	-4.848	1.5e-06	0.0014
Afatinib (rescreen)	KIN001-102	-0.325	-4.825	1.7e-06	0.0015
AICAR	Temsirolimus	-0.257	-4.817	1.8e-06	0.0015
17-AAG	FH535	0.344	4.813	1.8e-06	0.0015
BX-795	CCT018159	-0.238	-4.816	1.8e-06	0.0015
Bleomycin (50 uM)	PD-0332991	0.430	4.775	2.2e-06	0.0017
Embelin	WZ3105	-0.211	-4.769	2.3e-06	0.0017
Afatinib (rescreen)	PHA-665752	-0.313	-4.769	2.3e-06	0.0017
Gefitinib	GDC0941	-0.225	-4.761	2.4e-06	0.0018
BX-795	WZ3105	-0.261	-4.748	2.5e-06	0.0019
MLN4924	MS-275	0.374	4.753	2.6e-06	0.0019
Gefitinib	XL-880	-0.210	-4.732	2.7e-06	0.0019
Afatinib (rescreen)	AZD6482	-0.338	-4.729	2.7e-06	0.0019
Afatinib (rescreen)	OSI-027	-0.314	-4.727	2.8e-06	0.0019
Afatinib (rescreen)	BMS-754807	-0.305	-4.718	2.9e-06	0.0019
Olaparib	CCT018159	-0.187	-4.718	2.9e-06	0.0019
piperlongumine	XL-880	0.179	4.704	3.1e-06	0.002
Gefitinib	YM201636	-0.215	-4.699	3.2e-06	0.002
17-AAG	Rapamycin	0.343	4.649	4e-06	0.0024
Elesclomol	XL-880	0.340	4.652	4e-06	0.0024
MLN4924	Docetaxel	0.330	4.642	4.4e-06	0.0026
(5Z)-7-Oxozeaenol	GW 441756	0.251	4.621	4.6e-06	0.0027
Afatinib (rescreen)	PLX4720	-0.350	-4.610	4.8e-06	0.0027
Afatinib (rescreen)	PLX4720	-0.350	-4.610	4.8e-06	0.0027
Afatinib	GSK-1904529A	-0.289	-4.607	4.9e-06	0.0027
BMN-673	HG-6-64-1	0.315	4.604	4.9e-06	0.0027
Gefitinib	NVP-TAE684	-0.252	-4.605	5e-06	0.0027
Elesclomol	PHA-665752	0.346	4.602	5e-06	0.0027
Afatinib (rescreen)	GSK-1904529A	-0.308	-4.586	5.3e-06	0.0028
Gefitinib	AZD8055	-0.230	-4.588	5.4e-06	0.0028
Embelin	CCT018159	-0.186	-4.580	5.6e-06	0.0028
CX-5461	Pyrimethamine	-0.355	-4.572	5.7e-06	0.0028
Afatinib (rescreen)	GDC0941	-0.309	-4.570	5.8e-06	0.0028
Bleomycin (50 uM)	HG-6-64-1	0.332	4.568	5.8e-06	0.0028
XL-880	Nilotinib	0.235	4.564	5.9e-06	0.0028
Gefitinib	GSK-1904529A	-0.213	-4.563	6e-06	0.0028
T0901317	BIRB 0796	-0.191	-4.561	6e-06	0.0028
Afatinib (rescreen)	YM201636	-0.308	-4.552	6.3e-06	0.0029

With added covariate

Table C4: Drug sensitisation (pan-cancer with covariate)

Drug response	Sensitiser	Effect	Wald stat.	P-value	FDR
Afatinib	NVP-BEZ235	-0.397	-5.347	1.2e-07	0.00094
Afatinib	PI-103	-0.389	-5.344	1.3e-07	0.00094
BX-795	WZ3105	-0.299	-5.330	1.4e-07	0.00094
Afatinib	BMS-536924	-0.379	-5.230	2.3e-07	0.0012
AICAR	Temsirolimus	-0.268	-4.951	9.5e-07	0.0027
piperlongumine	XL-880	0.188	4.946	9.5e-07	0.0027
T0901317	BIRB 0796	-0.214	-4.935	1e-06	0.0027
Temsirolimus	AZD6482	0.407	4.923	1.1e-06	0.0027

Gefitinib	NVP-BEZ235	-0.263	-4.897	1.2e-06	0.0027
BX-795	CCT018159	-0.241	-4.887	1.3e-06	0.0027
17-AAG	FH535	0.343	4.799	2e-06	0.0037
Embelin	WZ3105	-0.212	-4.764	2.3e-06	0.004
Afatinib	Dasatinib	-0.339	-4.716	3e-06	0.0045
MLN4924	MS-275	0.371	4.719	3e-06	0.0045
17-AAG	Rapamycin	0.345	4.683	3.5e-06	0.0046
Afatinib	ZSTK474	-0.364	-4.674	3.6e-06	0.0046
Elesclomol	XL-880	0.354	4.628	4.5e-06	0.0053
Elesclomol	PHA-665752	0.371	4.620	4.6e-06	0.0053
Embelin	CCT018159	-0.186	-4.579	5.6e-06	0.0058
Afatinib	NVP-TAE684	-0.368	-4.570	5.8e-06	0.0058
NVP-BEZ235	AZD6482	0.340	4.571	5.8e-06	0.0058
MLN4924	Docetaxel	0.325	4.556	6.5e-06	0.0061
Afatinib	KIN001-102	-0.310	-4.523	7.3e-06	0.0065
piperlongumine	STF-62247	0.181	4.506	7.7e-06	0.0067
Embelin	CGP-60474	-0.194	-4.486	8.6e-06	0.0069
CCT018159	Rapamycin	0.209	4.478	8.8e-06	0.0069
EX-527	LY317615	-0.113	-4.462	9.5e-06	0.0069
Embelin	AT-7519	-0.191	-4.456	9.8e-06	0.0069
BX-795	Docetaxel	0.225	4.452	1e-05	0.0069
Gefitinib	PI-103	-0.233	-4.452	1e-05	0.0069
BX-795	CGP-60474	-0.235	-4.431	1.1e-05	0.0073
Elesclomol	Afatinib	0.320	4.370	1.4e-05	0.0093
Temsirolimus	XL-880	0.297	4.357	1.5e-05	0.0096
Afatinib	GSK2126458	-0.344	-4.340	1.7e-05	0.0099
AZD7762	PD-173074	0.249	4.336	1.7e-05	0.0099
ZM-447439	Docetaxel	0.226	4.325	1.8e-05	0.01
Afatinib	AZD8055	-0.311	-4.320	1.8e-05	0.01
Cytarabine	HG-6-64-1	0.273	4.312	1.9e-05	0.01
Vorinostat	Methotrexate	-0.205	-4.295	2e-05	0.011
Gefitinib	BMS-536924	-0.234	-4.275	2.2e-05	0.011
Gemcitabine	Docetaxel	0.470	4.275	2.2e-05	0.011
HG-6-64-1	Nilotinib	0.284	4.265	2.3e-05	0.011
Afatinib	BMS-754807	-0.262	-4.257	2.4e-05	0.011
Y-39983	PD-0332991	0.254	4.254	2.4e-05	0.011
HG-6-64-1	PHA-665752	0.311	4.254	2.4e-05	0.011
AZD7762	CCT018159	-0.254	-4.249	2.5e-05	0.011
Gefitinib	ZSTK474	-0.233	-4.229	2.7e-05	0.012
PD-173074	CCT018159	-0.187	-4.224	2.7e-05	0.012
XL-880	Nilotinib	0.229	4.210	2.9e-05	0.012
Gefitinib	Dasatinib	-0.234	-4.202	3e-05	0.012
Afatinib	AZD6482	-0.318	-4.193	3.1e-05	0.013
ZM-447439	CCT018159	-0.213	-4.177	3.4e-05	0.013
BX-795	Nilotinib	0.213	4.167	3.5e-05	0.013
Gefitinib	KIN001-102	-0.217	-4.160	3.6e-05	0.013
Afatinib	AP-24534	-0.299	-4.153	3.7e-05	0.013
YK 4-279	MS-275	0.260	4.156	3.7e-05	0.013
Tubastatin A	Pyrimethamine	-0.204	-4.150	3.7e-05	0.013
Camptothecin	HG-6-64-1	0.272	4.156	3.7e-05	0.013
piperlongumine	AZD6482	0.175	4.147	3.8e-05	0.013
Docetaxel	MS-275	0.263	4.139	4e-05	0.014
CCT018159	MS-275	0.202	4.133	4e-05	0.014
17-AAG	Masitinib	0.289	4.117	4.3e-05	0.014
Trametinib	Masitinib	0.370	4.116	4.3e-05	0.014
Embelin	HG-6-64-1	0.166	4.110	4.4e-05	0.014
YM155	GW 441756	0.432	4.105	4.5e-05	0.014
TW 37	GSK-1904529A	0.207	4.100	4.6e-05	0.014
piperlongumine	Rapamycin	0.168	4.077	5.1e-05	0.016
Vinblastine	ABT-888	0.260	4.077	5.1e-05	0.016
NVP-BEZ235	OSI-027	0.426	4.069	5.3e-05	0.016
17-AAG	HG-5-88-01	0.298	4.067	5.4e-05	0.016
Embelin	PF-562271	-0.182	-4.064	5.4e-05	0.016
Tubastatin A	GSK-1904529A	-0.211	-4.060	5.5e-05	0.016
PD-0325901	Y-39983	0.337	4.060	5.5e-05	0.016
Temsirolimus	Lapatinib	0.293	4.045	5.9e-05	0.016
Afatinib	GDC0941	-0.284	-4.042	5.9e-05	0.016

C.3 *Combination screen*

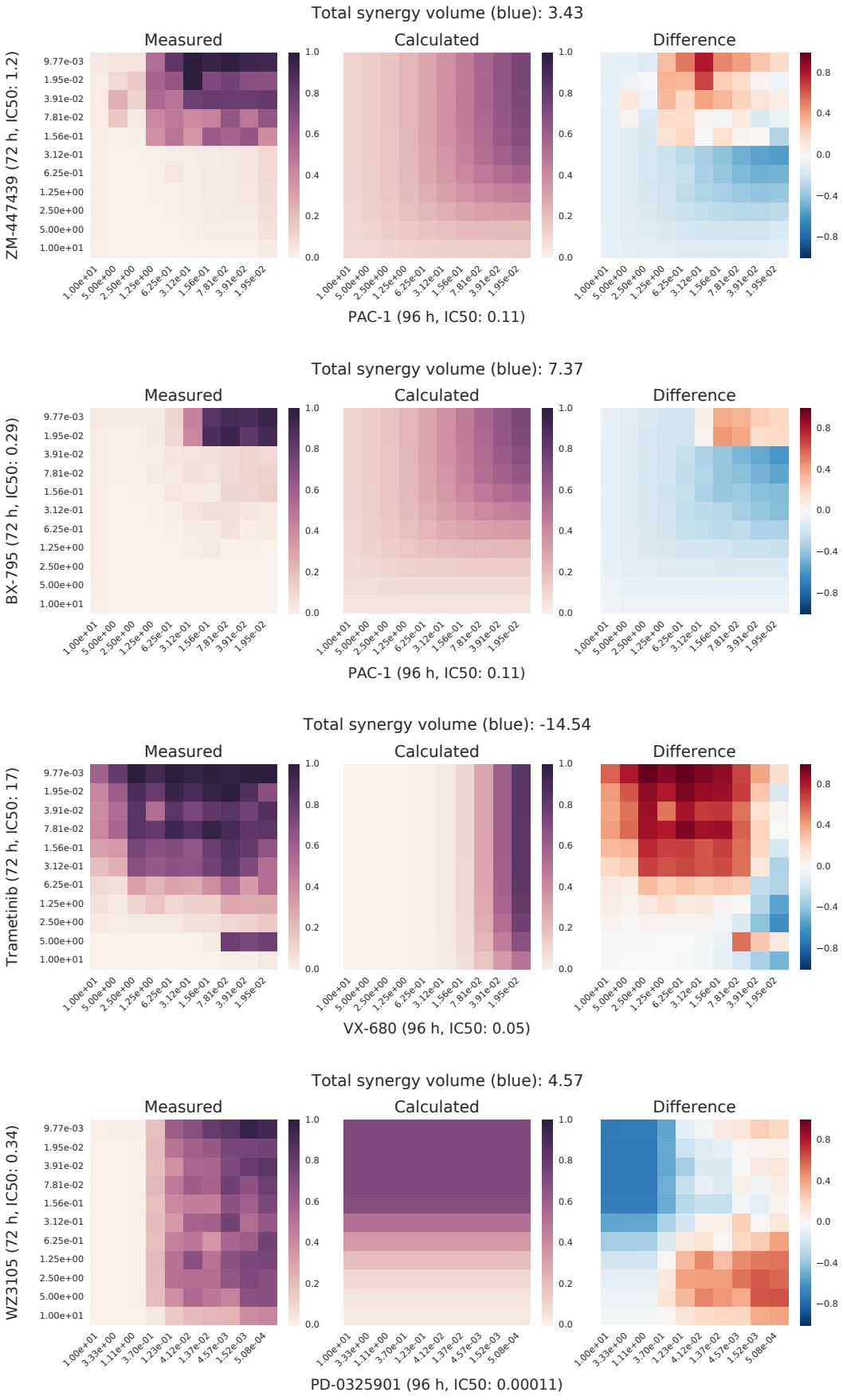

Figure C1: Predicted vs. measured synergy for drug combinations. Experimental result on the left, Loewe-additive model in the middle. Difference between the two on the right. Both axes represent micro-molar drug concentrations.

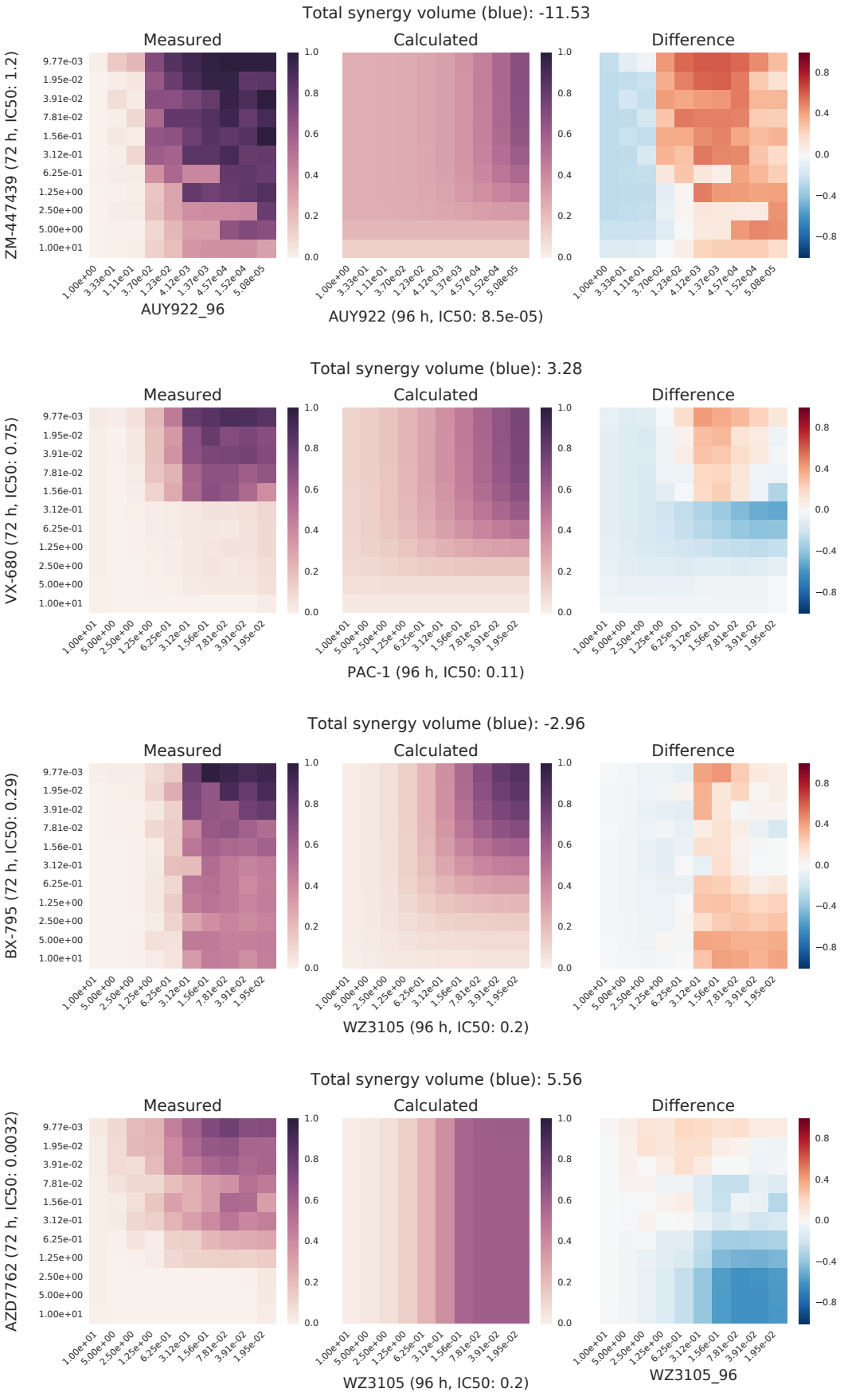

Figure C2: Figure C1 cont.

Figure C3: Figure C1 cont.

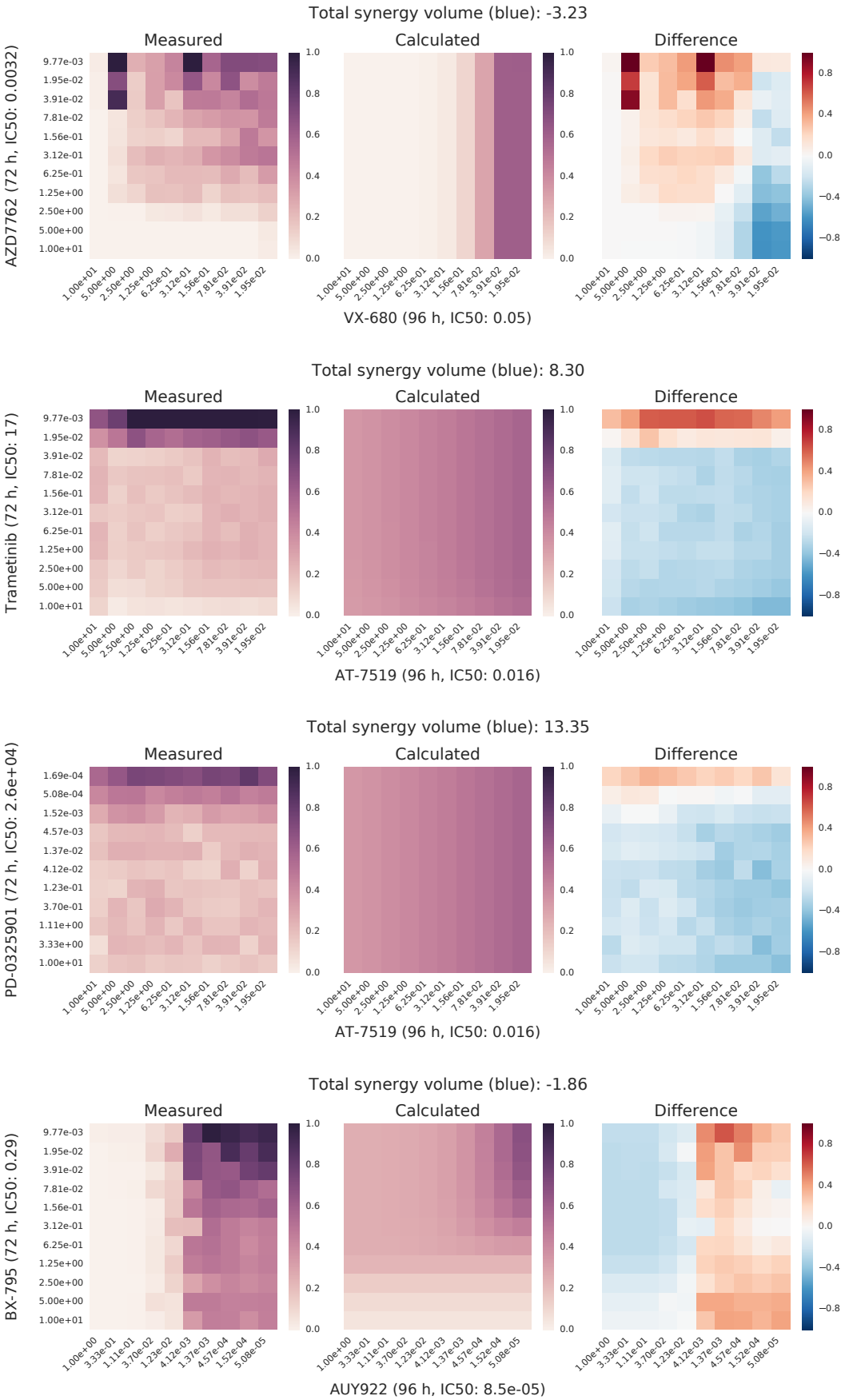

Figure C4: Figure C1 cont.