DETECTION OF KRAS SYNTHETIC LETHAL PARTNERS THROUGH INTEGRATION OF EXISTING RNAi SCREENS

Dissertation

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

von ELENI G. CHRISTODOULOU

geboren am 06 November 1982 in Athen/Griechenland

Betreuender Hochschullehrer: Prof. Dr. Andreas Beyer
Universität zu Köln
Prof. Dr. Michael Schroeder
Technische Universität Dresden

Dresden
Tag der Verteidigung: 15. 12. 2014

List of Publications

(1) Uddipta Biswas, Cornelia Wetzker, Julian Lange, Eleni G. Christodoulou, Michael Seifert, Andreas Beyer and Rolf Jessberger, "Meiotic cohesin SMC1β provides prophase I centromeric cohesion and is required for multiple synapsis-associated functions.", PLoS genetics, 9:12, 2013

(2) Weronika Sikora-Wohlfeld, Marit Ackermann, Eleni G. Christodoulou, Kalaimathy Singaravelu, and Andreas Beyer, "Assessing computational methods for transcription factor target gene identification based on ChIP-seq data.", PLoS computational biology, 9:11, 2013

(3) I. Tsamardinos, G. Borboudakis, E. G. Christodoulou and O. D. Røe, "Chemosensitivity Prediction of Tumours Based on Expression, miRNA, and Proteomics Data". In International Journal of Systems Biology and Biomedical Technologies (IJSBBT), 1:2, 2012

(4) E. G. Christodoulou, O. D. Røe, A. Folarin and I. Tsamardinos, "Information-Preserving Techniques Improve Chemosensitivity Prediction of Tumours Based on Expression Profiles". In Engineering Applications of Neural Networks, pages 453-462, Springer Boston, 2011

(5) Eleni G. Christodoulou, Vangelis Sakkalis, Vassilis Tsiaras, and Ioannis G. Tollis, "BrainNetVis: An Open-Access Tool to Effectively Quantify and Visualize Brain Networks," Computational Intelligence and Neuroscience, vol. 2011, Article ID 747290, 12 pages, 2011. doi:10.1155/2011/747290

(6) E. G. Christodoulou, M. Ioannou, M. Kafousi, E. Sanidas et. al. " A new gene expression signature related to breast cancer Estrogen Receptor status". In Proceedings of the 8th IEEE International Conference on BioInformatics and BioEngineering (BIBE 2008), October 8-10, 2008, Athens, Greece. IEEE Computer Society Press, pp 1-7

(7) E.G.Christodoulou, T. Dalamagas, T. Sellis, "Navimoz: Mining Navigational Patterns in Portal Catalogs", Current Trends in Database Technology-EDBT 2006, Springer, p. 801-813, ISBN 978-3-540-46788-5

The current thesis is a novel piece of work and does not share any findings with any of the above publications.

“As you set out for Ithaca hope your road is a long one, full of adventure, full of discovery...And if you find her poor, Ithaca won’t have fooled you. Wise as you will have become, so full of experience, you will have understood by then what these Ithacas mean.“

Konstantinos Kavafis

Abstract

KRAS is a gene that plays a very important role in the initiation and development of several types of cancer. In particular, 90% of human pancreatic cancers are due to KRAS mutations. KRAS is difficult to target directly and a promising therapeutic path is its indirect inactivation by targeting one of its Synthetic Lethal Partners (SLPs). A gene G is a Synthetic Lethal Partner of KRAS if the simultaneous perturbation of KRAS and G leads to cell death. In the past, efforts to identify KRAS SLPs with high-throughput RNAi screens have been performed. These studies have reported only few top-ranked SLPs. To our knowledge, these screens have never been considered in combination for further examination.

This thesis employs integrative analysis of the published screens, utilizing additional, independent data aiming at the detection of more robust therapeutic targets. To this aim, "RankSLP", a novel statistical analysis approach was implemented, which for the first time i) consistently integrates existing KRAS-specific RNAi screens, ii) consistently integrates and normalizes the results of various ranking methods, iii) evaluates its findings with the use of external data and iv) explores the effects of random data inclusion. This analysis was able to predict novel SLPs of KRAS and confirm some of the existing ones.

Acknowledgements

I first of all want to thank Michael Schroeder, for believing in me and accepting me in his group in the middle of my PhD. I want to thank him for his help and guidance in many crucial circumstances with which I came across. I also want to thank him for revealing my potentials to me by assigning tasks that I considered impossible but at the end I realized I could manage. I feel much more self confident now. Thank you so much Michael!

I definitely want to thank Andreas Beyer, the person who also believed in me at the very beginning and brought me to Dresden and to his group. He is the one who had the original idea and inspired me about this project. I thank him for all the scientific discussions that we had and all the things that he taught me. I wouldn't have managed equally well without the knowledge that he transmitted.

I want to thank Michael Kuhn the only person who followed my PhD from the beginning to the end, as he was also a member of Beyer group. He provided me with very useful help and suggestions on my project. I also want to thank George Tsatsaronis for his insightful discussions on evaluation using Precision-Recall. Zerrin and Janine also helped me a lot with their comments and hints, towards the completion of my work. I cannot forget Marit and Weronika who took my hand when I was fresh in my PhD and introduced me to their amazing projects and to the topics in which the group was involved. I want to thank Alex, our system administrator, for taking care of my workstation. He was always ready to help me and fixed my PC a lot of times. His crucial help, always so quick and efficient, accelerated a lot my work and reduced my stress! I also want to thank Claudia and Mandy for their help with administrative and bureaucratic matters that were a real struggle for me with my poor German. Many thanks to Mathiew, Betty, Susanne, Kalaimathy, Sinan, Johannes, Michael Seifert, Norhan, Alina, Maryllia, Jörg, Sebastian, Joachim, Simone, Daniel, Sainitin, Martin, Dirk, Christoph and Alicja for the wonderful atmosphere at work. Really thank you so much, I needed such a friendly surrounding to work efficiently and to forget many of my problems while at work...

I also want to convey my thanks to Ji Luo, David Hancock and Julian Downward for helping me with data and information on the screens they conducted and which are the basis of this analysis, as well as Jaiswal Alok too for providing me information on the Wang screen. I want to thank Joe Tym, the correspondig person for CanSar database, who provided me with the Barbie screen. I thank Irene Kaplow, Joshua Gould and Yingyao Zhou for their feedback on RNAiCut, RIGER and RSA respectively.

I want to thank my dear friends Dimitra, Hannes, Carine and George for their invaluable friendship and for their support during a very difficult period. Many thanks to sweet Norhan for standing by me and for being a great host!

I finally want to thank the most important people in my life; my parents George and Georgia, my sister Maria and my fiance Stelios. I deeply thank them for making me what I am and for supporting me in the difficulties that I faced during my PhD. Special thanks to Stelios for inspiring and encouraging me to proceed a PhD and for standing by me all the way through.

Contents

List of Publications i

Abstract iii

Acknowledgements iv

Contents vi

List of Figures ix

List of Tables xi

Abbreviations xii

1 Motivation 1
- 1.1 The open problems 2
  - 1.1.1 Clinically relevant problem: Treatment of KRAS dependent cancers remains unresolved 2
  - 1.1.2 Studies do not agree: Why can computer science help us to this aim? 3
  - 1.1.3 Problems with integration 3
  - 1.1.4 Comparison of Rankings and General Applicability 3
- 1.2 Strategy 4
- 1.3 Thesis Outline 4

2 Background 6
- 2.1 Cancer and KRAS 6
  - 2.1.1 GTPase KRAS 7
  - 2.1.2 Clinical significance 8
- 2.2 Synthetic Lethality 9
  - 2.2.1 Synthetic Lethality for Drug Discovery 9
  - 2.2.2 RNAi screening as a tool for SLP discovery 11
  - 2.2.3 OTEs and noise in RNAi screens 12
  - 2.2.4 Cell Lines used as Tumour Models 14
- 2.3 Background in aggregation of ranks 14

3 Data and Methods 17
3.1 The datasets . . . 17
3.2 Assignment of a single value in the case of multiple shRNAs per gene. . . 20
3.2.1 Standard method . . . 20
3.2.2 RIGER . . . 22
3.2.3 RSA . . . 24
3.3 RNAiCut . . . 25
3.4 Rank Aggregation . . . 26
3.4.1 Sophisticated Rank Aggregation using heuristics . . . 26
3.5 Random or additional screens: How much do they alter the result? . . . 27
3.5.1 Extra screen: Barbie . . . 28
3.5.2 Random screen . . . 29

4 Results 30
4.1 KRAS SLPs in Literature . . . 33
4.1.1 KRAS SLPs from RNAi screens . . . 35
4.1.2 KRAS SLPs from the rest of literature . . . 35
4.1.3 Overlap . . . 36
4.2 Getting to grips with the datasets . . . 36
4.2.1 Multiple shRNAs per gene . . . 37
4.2.2 How differences in knock-down technologies and readout methods were encountered? . . . 38
4.3 How was screen size variance encountered? . . . 39
4.4 How were cell line differences encountered? . . . 40
4.5 Retrieved genes from the three main methods: Standard, RIGER and RSA . . . 41
4.6 Genes with many of their shRNAs ranking below threshold . . . 45
4.7 Results from RNAiCut . . . 47
4.8 Incorporation of the genes retrieved by rank aggregation . . . 48
4.9 First hit genes selection and pathways enrichment . . . 50
4.10 Comparison with literature genes . . . 50
4.11 Comparison with high throughput drug screens . . . 52
4.11.1 Dataset retrieval . . . 54
4.11.2 CCLE, GDSC and NCI60 screens . . . 55
4.11.3 Steckel Drug screen . . . 56
4.11.4 Drug-Target matching . . . 56
4.11.5 Identified genes from the analysis of drug response data . . . 59
4.12 Network enrichment . . . 60
4.12.0.1 Pathway Commons analysis . . . 61
4.12.0.2 Gene Ontology analysis . . . 62
4.12.0.3 Enrichment having as basis human kinases . . . 63
4.12.0.4 Discussion on network enrichment . . . 66
4.13 Final selection of genes for experimental validation . . . 67
4.14 Analysis and result on the incorporation of an extra or random screen . . 70

5 Discussion **72**
5.1 Final selected genes and their roles in cancer development . . . 72
5.2 Overlap of Gold Standard Genes with RankSLP findings . . . 76
5.3 Overview of methods used in this thesis . . . 76
5.3.1 Three basic methods for hit detection . . . 76
5.3.2 Evaluation of Standard, RSA and RIGER methods: Which performs best? . . . 78
5.3.3 Rank aggregation methods . . . 80
5.4 Assessment of screen quality . . . 82
5.5 How does the selection of thresholds and methods influence the final SLPs? . . . 83
5.6 Connection of this work with previous methods and findings . . . 86

6 Conclusions and future extensions **88**
6.1 Summary . . . 88
6.2 Complexity . . . 89
6.3 Contributions . . . 91
6.4 Possible extensions . . . 93

A Appendix **96**
A.1 siRNAs and shRNAs . . . 96

Bibliography **99**

List of Figures

1.1 Motivation . . . 2

2.1 Oncogene . . . 7
2.2 KRAS function . . . 8
2.3 Synthetic Lethality . . . 9
2.4 RNAi screening . . . 12
2.5 Rank Aggregation: Illustration . . . 15
2.6 Spearman's footrue-Kendall's tau . . . 16

3.1 Dropout screens using barcodes. . . . 19
3.2 Standard Method Ranking . . . 21
3.3 RSA algorithm . . . 24

4.1 Approach with schematic 1 . . . 31
4.2 Approach with schematic 2 . . . 32
4.3 normalized datasets: The viability values of each replicate . . . 37
4.4 Wang screen: shRNA values differ . . . 38
4.5 Wang screen: Differences in microarray fluorescence . . . 39
4.7 Structure of wild type and mutant KRAS . . . 41
4.8 Scatterplot of Standard method, RSA and RIGER results . . . 43
4.9 Intersections of top shRNAs 1 . . . 45
4.10 Intersections of top shRNAs 2 . . . 45
4.11 Intersections of top shRNAs 3 . . . 46
4.12 Weights for the rank aggregation . . . 49
4.13 Consensus ranks. GSG retrieval . . . 52
4.14 RIGER-RSA and GSG ranking . . . 53
4.15 Drug basics . . . 54
4.16 Statistical analysis of datasets . . . 55
4.17 STITCH graph . . . 57
4.18 STITCH target frequency vs confidence threshold . . . 58
4.19 Pathway commons analysis: TNF alphaNK-kB . . . 62
4.20 Pathway commons analysis: 9 pahtways . . . 63
4.21 GO: Positive regulation of macromolecule metabolic process . . . 64
4.22 GO: (Transcription) regulatory region DNA binding . . . 65
4.23 Enrichment having as basis the human kinases . . . 67

5.1 Hallmarks of cancer and the hit genes . . . 73
5.2 Precision-Recall curves for Standard method, RIGER, and RSA . . . 79
5.3 Baseline rank aggregation methods . . . 80

5.4 Precision-Recall curves for Rank Aggregation methods 81

6.1 Polypharmacology 93

A.1 RNA interference 96
A.2 Dicer function 97

List of Tables

3.1 RNAi datasetes 18
3.2 Random or extra screen 28

4.1 KRAS SLPs in literature 33
4.2 Retrieved genes by each method. Intersection among the 3 screens. 44
4.3 Significance of Intersections 47
4.4 RNAiCut *vs.* Standard method findings 47
4.5 RNAiCut *vs.* all main methods' findings 48
4.6 RankAggreg *vs.* all mainmethods 50
4.7 Pathways in which the hit genes participate 51
4.8 Chemical Genomics Screens 53
4.9 Drug screen analysis results 59
4.10 CCLE: Intended targets of significant compounds 59
4.11 Enriched GO terms with respective p-values 66
4.12 Sources of evidence for each gene 69
4.13 Random/Barbie screen inclusion findings 70

5.1 Assessment of methods and thresholds chosen 83

Abbreviations

BRCA1/2	**BR**east **CA**ncer **1/2**
CNV	**C**opy **N**umber **V**ariation
GAP	**G**TPase **A**ctivating **P**roteins
GDP	**G**uanosine **D**i**P**hosphate
GSG	**G**old **S**tandard **G**enes
GTP	**G**uanosine **T**ri**P**hosphate
LOF	**L**oss **O**f **F**unction
mut	**mut**ant
PARP	**P**oly **A**DP-**R**ibose **P**olymerase
PPI	**P**rotein **P**rotein **I**nteraction
RNAi	**RNA** **I**nterference
shRNA	**s**hort **h**airpin **RNA**
siRNA	**s**mall **i**nterfering **RNA**
SLP	**S**ynthetic **L**ethal **P**artner
SNP	**S**ingle **N**ucleotide **P**olymorphism
wt	**w**ild **t**ype

Dedicated to my parents George and Georgia,
and to my beloved Stelios

Chapter 1

Motivation

In the past, various synthetic lethal screens have been performed aiming at the detection of KRAS SLPs, but their results agree only partially, as extensively presented in section 4.1. The problem is that due to this inconsistency and due to improper cell line selection (see subsection 2.2.4), almost none of the detected SLPs was really successful in clinical trials. The only exception is bortezomib (velcade), a compound targeting the APC/C complex. Bortezomib was approved in 2003 by FDA for use in lung cancer patients with the KRAS mutation [1]. However, since 2003, many other studies concerned with KRAS SLPs have been driven. The aim is to find other drugs that target KRAS SLPs and can effectively kill other than lung cancer tumours as well.

Until now, every new effort towards the detection of KRAS SLPs has been involving the creation of a new biological dataset, comprising of KRAS mutant and KRAS wild type cells. To our knowledge, only one study based on integration of existing datasets has been performed [2], despite the invaluable information that they may hide. This is mainly due to the many challenges that such an integration underlies; within each new experiment different cell lines are screened, having variable KRAS mutations and using different RNAi libraries (i.e. this means that different genes are targeted in each study). On top of this, the inherent noise of RNAi screens, owed primarily to OTEs (see section 2.2.3), has transformed the dataset integration to an undoubtedly difficult problem. The current thesis is concerned with a computational integration analysis of existing RNAi datasets for the prediction of new KRAS Synthetic Lethal Partners (SLPs). The aim is to find a consistent hit that ranks *relatively high* among all screens. Therefore, it is more robust and more reliable as real KRAS SLP, and could be applied to a new tumor. To this end, a novel strategy which involves a statistical analysis approach has been developed. The current implementation allows for a completely new perspective on the existing datasets. This is of course facilitated and driven by the nature of RNAi screens; their advantage is exactly the fact that they are performed in large scale. These high-throughput datasets constitute a perfect input for statistical and computational analysis, which can lead to novel findings in a quick and cheap way.

This first step in integrative analysis of RNAi data may inspire the scientific community towards the direction of their consistent integration. The motivation of this thesis is summarized in figure 1.1.

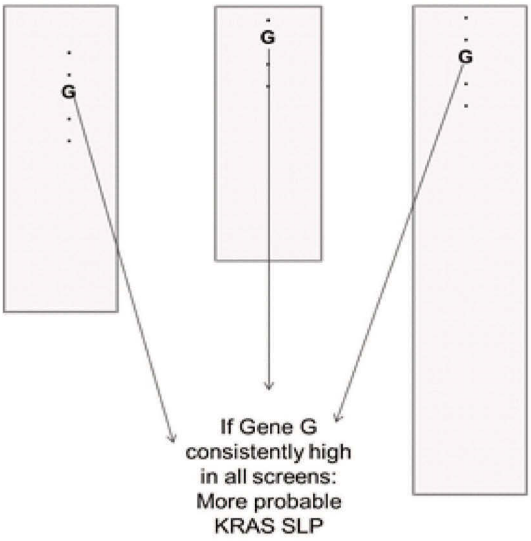

G
G
G
If Gene G consistently high in all screens: More probable KRAS SLP

FIGURE 1.1: If a gene ranks consistently high in all screens, this provides more evidence in favor of the detected drug target functioning in many patients.

1.1 The open problems

1.1.1 Clinically relevant problem: Treatment of KRAS dependent cancers remains unresolved

The open problem from the medical point of view is that KRAS dependent cancers haven't met an efficient means of treatment yet. Many synthetic lethal partners have been suggested but very few of them are confirmed by followup experiments, making difficult to treat clinical cases of KRAS mutant cancers. With the increasing rates of KRAS incidence in human cancers, the need for a treatment becomes more and more imperative. The computational approach followed in this thesis by integrating multiple datasets and methods, although still far away from the clinics, will provide robust resulting genes that can act as targets towards the treatment of these cancers and might convince biologists and physicians to conduct wet lab experiments accordingly.

1.1.2 Studies do not agree: Why can computer science help us to this aim?

A number of biological RNAi studies towards detection of KRAS SLPs have been conducted. There should be some additional knowledge in the already existing data and an effective way to extract it from all relevant information needs to be found. The provided datasets are high-throughput, high dimensional and involve a lot of varying parameters. Thus computer science techniques are a proper path to follow, without the need to perform additional experiments, that require invaluable time and money. Data integration towards KRAS SLP detection is something that, to our knowledge, hasn't been addressed in the past. To our knowledge, no bioinformatics methods specific to the detection of synthetic lethal pairs have been developed yet. It is a problem that remains open and multiple ways of addressing it will be covered in this work. However there is one study that follows an integrative approach using some of the methods that RankSLP is also using, towards the detection of genes affecting sensitivity to tamoxifen [2]. This study doesn't compare its finding with external data though, which is the innovative part in this thesis.

1.1.3 Problems with integration

Consistent RNAi screening data integration is not a trivial task as there are many parameters that vary among the screens. Some of them are differing screen sizes, different cell lines e.t.c. This inconsistency was a big challenge to this analysis. Appropriate selection of datasets and computational methodologies in order to bring the data in a comparable format was investigated in this thesis.

1.1.4 Comparison of Rankings and General Applicability

This thesis examines how different ranking methods relate, employing the KRAS example. Calculation of rankings by employment of several methods, and aggregation of the ranking results are of interest and of general applicability in many fields. From a common example, like the selection of a University to proceed one's studies, to a more specific but old case, like the design of a robust voting scheme, efficient rank aggregation is present in many aspects of human life.

1.2 Strategy

To alleviate the open problems, different ranking methods were compared and their findings were integrated. The established methodology is divided in two parts:

- **RNAi screen ranking.** This part involves the application of already developed methods to prioritizing the hits of an RNAi screen (RIGER, RSA, RNAiCut e.t.c.). The parameters of these methods were tuned to meet the needs of the provided datasets. Multiple shRNAs per gene are effectively treated both by these existing methods and by a newly developed approach, based on the selection hit frequency of an shRNA (Venn diagrams - see section 4.6). This first part also includes the aggregation of rankings from the applied methods to conclude with a first set of potential candidates.
- **Evaluation of findings.** This second part makes use of external datasets, like networks, compound inhibition screens and existing literature towards filtering the results of the first part. Proper ranking techniques are combined with network enrichment of the initial hit genes. At the end, the similar genes between both parts are collected and the ones that don't agree (false positives), probably due to OTEs, are discarded.

1.3 Thesis Outline

The remainder of this thesis is organized as follows:

- Chapter 2 initially provides insight into the biology that underlies wild type and mutant KRAS function. It then presents synthetic lethality as a way of treating KRAS mutant tumours. The basic principles of RNAi screens are explained, which are the type of datasets that are used in the current work. Chapter 2 concludes by adding a background on aggregation of ranks, which is one of the main topics of this thesis.
- In Chapter 3, at first an overview of the RNAi datasets used in this thesis is given. The different ranking algorithms that are applied on these data in order to prioritize the hit genes are described. The techniques and algorithms that are implemented are extensively presented.
- In Chapter 4, the known from literature KRAS SLPs are reported. It is shown that existing findings do not agree and the importance of detecting KRAS SLPs that are

consistent among various datasets is stressed. This Chapter further discusses the problems and considerations on RNAi dataset integration and RankSLP's methodology for overcoming them. The results of applying the methods described in Chapter 3 are presented. Furthermore, several ways of evaluating the findings are presented: Comparison with SLPs from literature and with external chemical genomics screens, along with integration with gene network information. Creation of a randomized dataset and selection of genes that are immune to noise are also performed. Chapter 4 concludes with a number of more probable KRAS SLPs identified by the current analysis.

- In Chapter 5, the results are discussed and reasonable explanations to support them are provided. The ranking methods used by RankSLP are compared and evaluated. The advantages of the developed approach along with potential pitfalls are also discussed.

- Finally, Chapter 6 summarizes the findings of this thesis and highlights how the open problems that are mentioned in the Motivation were solved. The contribution of this thesis to cancer systems biology and to computer science is shown. Potential future expansions are also mentioned.

Chapter 2

Background

Chapter 2 provides a broad introduction on KRAS and its biological function when in the mutant and when in the wild type state. Synthetic lethality is explained and suggested as a promising approach to the treatment of KRAS mutant tumours. Since in this thesis synthetic lethal KRAS-specific RNAi screens will be used as basis, the basic principles of RNAi screening are explained here. Potential problems with RNAi screens and reasons for disagreement with clinical studies are discussed. This Chapter concludes by adding a background on aggregation of ranks, which is one of the main topics of this thesis.

2.1 Cancer and KRAS

It is now well known that, in eukaryotic cells, DNA undergoes continuous modifications during an individual's life, including damage. Cells have different strategies for counterbalancing DNA damage through DNA repair pathways. These pathways work properly in normal cells. However, in cells where mutations in some crucial genes are accumulated, the effectiveness of these pathways may be disrupted, which in turn transforms these cells to cancer cells.

Mutations are crucial, not only for the initiation of cancer but also for the support of the cell's tumorigenic state and for cancer progression [3]. The three main types of genes that play a role in cancer are oncogenes, tumor suppressor genes and stability genes [4]. Our gene of interest, KRAS, is a proto-oncogene. A proto-oncogene can become oncogene due to elevated gene expression levels, due to chromosomal translocation or due to mutations within either the proto-oncogene itself or one of its regulatory regions. In the case of KRAS and of the focus of this thesis, an activating mutation in just one of

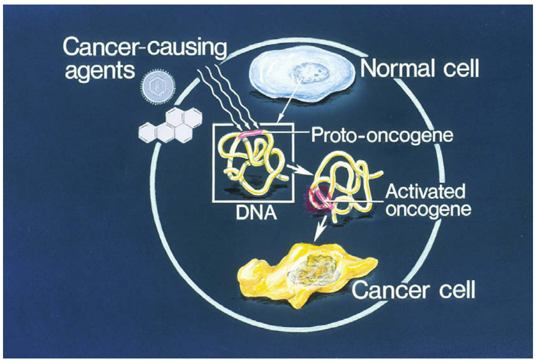


FIGURE 2.1: This graphic illustrates the stages of how a normal cell is converted to a cancer cell, when an oncogene becomes activated. Figure and caption are provided by National Cancer Institute, having AV Number: AV-8808-3615.

its alleles can transform it to oncogene and suffices to enhance tumorigenesis on the cell, since the presence of an oncogene makes a cell more susceptible to cancer. The term "oncogene" is attributed to the National Cancer Institute scientists who aptly represent with figure 2.1 its connection to cancer.

2.1.1 GTPase KRAS

KRAS stands for V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog, a name that stems from its first discovery in rats. KRAS is a member of the tyrosine kinase pathway and it is a small GTPase molecule: In its wild type state it acts as a GTP to GDP converter, facilitated by the GAP protein. The function of KRAS is presented visually in figure 2.2.

When KRAS is in its *active* or *GTP*-bound state, it contributes to the propagation of growth factor signals from the extracellular environment to the nucleus. Growth factors are responsible for stimulating important to the cell processes, such as cellular growth, proliferation, healing, and cellular differentiation.

Normally, KRAS is inactivated again by GAP. However, mutated KRAS loses this capability and remains locked in its active state. As a result, when KRAS is mutated, the MEK/ERK pathway, which is a proliferation pathway, is constitutively activated. The proliferation signals are continuously propagated leading to unnecessary growth and subsequently to tumour formation.

An important point is that a KRAS mutant cell can be KRAS addicted or not, meaning that its growth and survival can often be impaired by the inactivation of the KRAS oncogene alone. This phenomenon is often observed in cancer cells and is referred to as 'oncogene addiction'. Interestingly, the KRAS addiction of a cell is influenced

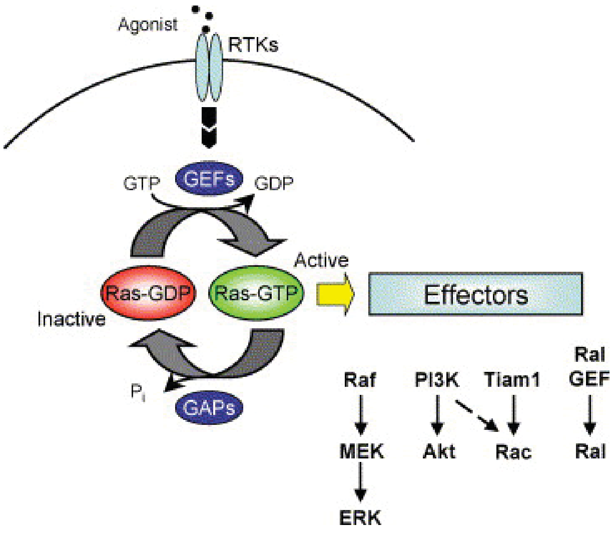


Figure 2.2: Figure taken from [5]. Wild type and mutant KRAS function and the effector pathways that are constitutively activated when KRAS is locked in its GTP-bound state.

by epithelial - mesenchymal transition (EMT), a procedure in which epithelial cells lose their cell polarity and cell-cell adhesion, acquire invasive properties and finally get transformed to mesenchymal stem cells. Epithelial KRAS-mutant cells usually being KRAS-dependent and mesenchymal cells usually being KRAS-independent [6]. The same study underlines also that a cell's dependency on KRAS increases with KRAS amplification.

2.1.2 Clinical significance

Pancreatic cancer is an especially therapy resistant form of cancer. Some very famous personalities, like Steve Jobs, Luciano Pavarotti and Patrick Swayze died of it. KRAS is mutated in 90% of pancreatic cancers [7]. Moreover, somatic KRAS mutations are very often detected in leukemias, ovarian cancer, colon cancer, thyroid cancer and lung cancer as well [8]. As long as all cancer types are considered, KRAS is one of the most activated oncogenes, with 17 to 25% of all human tumors harboring an activating KRAS mutation. A single nucleotide substitution is sufficient to turn KRAS on and lock it to its GTP-bound state. The critical activating KRAS mutations happen in codons 12, 13, 59, 61 and 63. Most frequent among them is the mutation in codon 12, followed by the mutations in codons 13 and 61 [8]

2.2 Synthetic Lethality

Synthetic lethality is a genetic phenomenon which was first observed in 1922 in Drosophila melanogaster [9]. The term, however, was generated by Dobzhansky in 1946, to describe complementary lethal systems in wild-type population of Drosophila pseudoobscura. He induced synthetic lethality by recombining two homologous chromosomes of different origin that had been perfectly viable as homozygotes. "Certain genes, born by each original chromosome, were now on the same chromosome and interacted to produce a recessive lethal effect" [10]. Synthetic lethality happens when the simultaneous perturbation (including mutation) of two genes results in cellular or organismal death, whereas individual perturbation of one of them doesn't lead to lethal phenotype. The concept of synthetic lethality is depicted in figure 2.3.

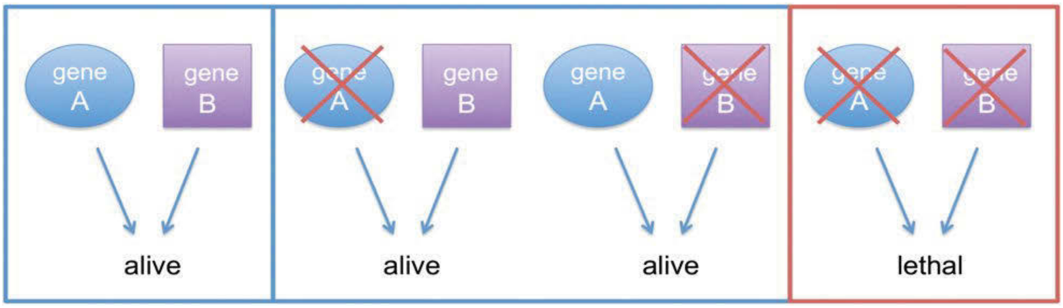


FIGURE 2.3: Schematic representation of synthetic lethality. Two genes are synthetic lethal only when their simultaneous inactivation results in cellular or organismal death. In this example, perturbation of either gene A or gene B does not affect viability whereas perturbation of both at the same time is lethal. [11]

2.2.1 Synthetic Lethality for Drug Discovery

The majority of chemotherapeutic drugs are able to kill rapidly growing cells. However, many of these drugs also kill healthy cells along with cancer cells [12]. Some examples are doxorubicin (toxic to the heart), bleomycin (toxic to the lung) and cytarabine (toxic to the cerebellum) [13]. Another undesired outcome of current cancer treatment is the development of chemotherapy resistance in tumors [14]. So, here comes the challenge for the drug industry: The need to develop highly selective drugs, with reduced side effects, is urgent. Thus, more recent cancer drug discovery is aiming at developing selective drugs adapted to the characteristics of the specific tumor (personalized therapy). The concept of synthetic lethality seems to be promising towards this aim, as it may explain the sensitivity of cancer cells to certain drugs. Cancer cells have different properties than normal cells, and their gene and protein networks expose a different *wiring* [15].

Conventional treatment approaches are usually non effective, because they are not based on these special properties of cancer cells and are often as general as to kill normal cells as well. New approaches, as synthetic lethality, are better suited for the gene interactions in cancer cells. Current drug development efforts are being shifted to this direction. Synthetic lethality takes advantage of the high instability of tumor cells and can be addressed either in cases of loss-of function mutants or in other types of perturbations, like gene over-expression, environmental influence (e.g. stress) and others. This high instability is a unique property of the cancer cells and should be exploited. As Paul Workman, Ph.D., director of Cancer Research U.K.'s Center for Cancer Therapeutics, said, "What do cancer cells have that normal cells don't?... They have mutations, and you can take advantage of those" [16].

Therapeutic strategies leveraging synthetic lethality have recently been brought to clinical trials having very encouraging first results [17]. The most promising example is the synthetic lethal interaction of PARP and BRCA1/BRCA2 in cases of ovarian and breast cancer, which was very successful in phase II clinical trials [18]. The gene PARP is essential for the repair of single-strand DNA breaks in cases of DNA damage. However, in cases where BRCA gene is mutated, the cell has become a cancer cell and PARP, by fixing the DNA breaks helps to the survival of this cancer cell. When PARP action is inhibited, and at the same time both copies of the BRCA gene are mutated, the cancer cell is led to apoptosis.

A second example is the treatment of VHL mutant renal cancer by inhibiting its SLP mTOR, which made it to phase II clinical trials [7]. Moreover, some treatments based on SLP pairs gave encouraging results in phase I trials, like MSH2 + methotrexate [7], and Chk1 /2 + carboplatin or cisplatin [19]. These are now embarking for phase II trials and there is high confidence that they will be successful.

Until recently, KRAS was considered *undruggable*, meaning that it cannot be directly targeted by small molecules [20, 21]. The reasons for that are its high binding affinity for GTP/GDP (picomolar range) and the lack of knowledge for other allosteric regulatory sites [22]. However, in 2013, Ostrem *et* al. [23] managed to develop small molecules which are able to irreversibly bind to KRAS G12C mutant and de-activate it, without influencing the wild type protein. The result of this binding is the disruption of both switch-I and switch-II sites of mutant KRAS, preventing KRAS from *locking* to the active GTP-state, and thus preventing the initiation of downstream tumorigenic pathways (i.e. MEK/ERK pathway) Until the work [23] was published, where a direct way to target KRAS was detected, efforts had been shifted to targeting KRAS *indirectly.* The concept of synthetic lethality seems to be promising towards this aim [13], because it is a more stable way of impairing the tumorigenic action of KRAS and also covers the cases where the rest of the codons (apart from codon 12) are mutated.

A number of RNAi screens aiming at the identification of such synthetic lethal partners of KRAS have been conducted [6, 24–28].

2.2.2 RNAi screening as a tool for SLP discovery

RNAi screens are a tool for gene *knock-down*, not *knock-out*. This means that they are used to destroy the mRNA that is transcribed by a gene, without affecting the respective gene in the cell's DNA. This is currently achieved through either synthetic siRNAs or vector-expressed short hairpin RNAs (shRNAs), having complementary sequence to the mRNA produced by the gene of interest. Please consult the Appendix for a deeper insight into shRNA and siRNA function and differences. RNAi should ideally be very selective and knock out its target without affecting other genes. Its major property is its simplicity; it is much more simple than modifying a gene. This property made RNAi technology a considerable player in the treatment of cancer, since it could ideally target the cancer-causing gene in a simple way [29].

RNAi technology soon enabled high throughput screens in cell cultures. The respective small and large scale RNAi action is depicted (see figure 2.4).

Some highlights regarding RNAi screens are:

- One of their largest contributions is their use as *Loss-Of-Function (LOF)* screens for drug discovery [31, 32]. In this context, they enable gene characterization with respect to their LOF phenotypes. This is a simple tool towards gene discovery, with the potential of being an interesting drug target.
- Another very interesting application is their use as *modifier* or *pathway* screens, where 'RNAi is used to identify genes and pathways that, when silenced, can either enhance or suppress a given phenotype of interest' [31].
- A specific category of the *pathway* screens, which can be considered as a stand-alone category, are *synthetic lethal* screens. In this case, various genes are silenced in pairs, and a lethal to the cell combination is sought. In the setting of cancer drug discovery, we are interested in cancer cells that contain a mutant allele of a gene (perturbed from its normal function) and the gene which, when knocked-down leads to cell lethality, is sought. The first who introduced RNAi screening towards the detection of synthetic lethal partners of cancerous gene mutants was Hartwell in 1997 [33].

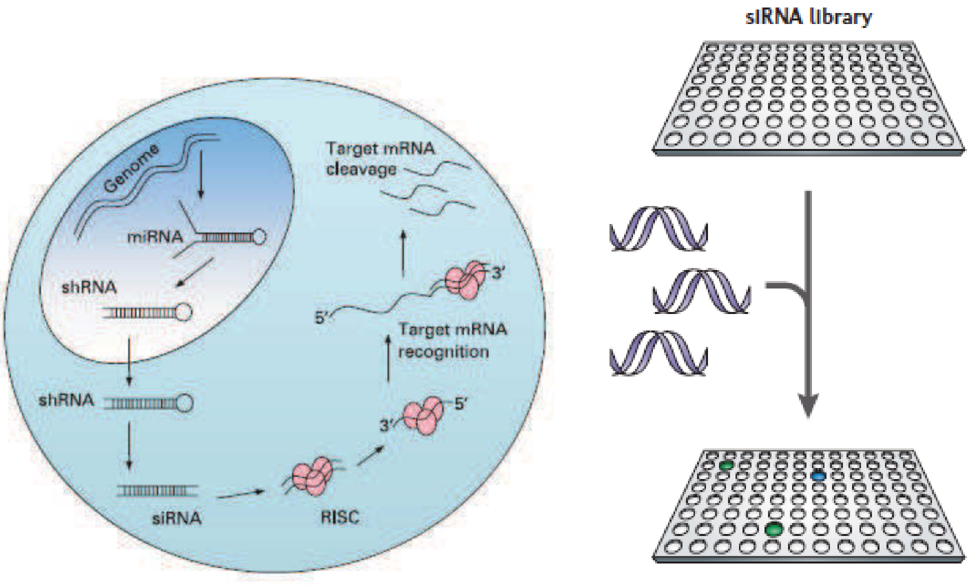


FIGURE 2.4: Left: Figure and caption adapted from [30]. Low scale intracellular process of RNAi interference function. For simplicity we have shown how microRNAs (miRNAs) can mediate RNA interference in mammalian cells by causing the degradation of protein-coding transcripts. What usually happens in experiments is that viruses or plasmids containing miRNA-coding or shRNA-coding sequences are introduced into mammalian cells and these mimic the production of endogenous miRNA and shRNA and are processed into siRNA in the same way. Alternatively, the siRNA can be directly inserted in the cell. Right: Figure and caption adapted from [31]. Large scale, high-throughput RNAi screening. In standard siRNA transfection, siRNA is added to pre-plated cells.

It can thus be concluded that RNAi screens are a tool for synthetic lethal pair detection. This is also clearly stated by William G. Kaelin in [13]. However, the high genomic redundancy of human cells can make the discovery of novel and significant genetic interactions very difficult. It is also a time-consuming and expensive procedure [12].

2.2.3 OTEs and noise in RNAi screens

Since 2006, when Fire and Mello were awarded the Nobel Prize in Physiology or Medicine for detecting that one can suppress a target gene's expression using RNA interference (RNAi), RNAi screening has been extensively used for drug target detection [34, 35]. Although many experiments have been performed, only a handful of the top-identified genes proceeded to follow-up experiments. Astonishingly, none of them was verified at second stage [36]. This threw down the gauntlet for the scientific community which started seeking the reason of this discrepancy. One possible explanation is that the observed effects on the cell of interest are the result of an interaction / synergy between siRNAs or shRNAs and other cellular components or genes. This means that the observed phenotype (i.e. lethality) is due to combinatorial effects and may not be target-specific. Each oligo siRNA or shRNA sequence, exactly because of its small size, can be complementary to many other genes apart from the target gene. One possible reason for these unintentional binding is that DICER, the enzyme that cleaves the double-stranded precursor miRNA to create miRNA, from which later on siRNA is synthesized, is not always accurate [37]. A recent paper found that DICER recognizes also the 5' RNA end [38], and not only the 3' end as it was believed in the past. So, mutations in the 5' end may affect DICER's cleavage ability.

That is why the observed signatures are often *"siRNA specific rather than target specific"* [39]. In [39] the authors also suggested pooling many siRNAs together for better result. In brief, this alerts us that much of the observed phenotype is random effects or *noise*. For RNAi screening this term is called "Off Target Effects" (OTEs). A recent publication [40] discusses the high prevalence of OTEs, which are sometimes even more than the intended, "on-target" effects of an RNAi screen. The authors performed quantitative analysis on 3 RNAi published datasets and discovered that there exist some "seed" sequences that systematically block cellular infection by pathogens, independent of the intended effect. They suggest the design of novel siRNA oligo sequences, and the production of new libraries, that contain deliberately a number of such seeds. Biological experiments, RNAi screening included, should always be performed in replicates. In the case of the new suggested libraries, with repetitive experiments, the seeds will be repetitively screened. Thus, their effects will be systematic in every repetition, and in this way they will be more easily detected. Consequently, this will allow screen performers to correct for the bias they introduce.

In summary, OTEs can influence a lot the outcome of an RNAi screen. This does not leave the KRAS synthetic lethal RNAi screens unaffected. Previous KRAS -directed RNAi studies came up with genes that were not further validated. For example, Scholl et *al.* [28], identified STK33 as KRAS SLP. Follow-up studies however failed to observe a significant differential viability effect between knocking down STK33 in KRAs mut and KRAS wt cells [41, 42]. Barbie et *al.* [24], reported TBK1 as KRAS SLP. Once more, further studies did not confirm this finding [43]. A final example is PLK1, identified by Luo et. *al.* [25] as KRAS SLP. However, there are many studies on different cancer types from other four groups, which expose the general role of this gene in cancer [44, 45] and in cellular viability [46–48]. This perhaps implies that the observed effect by Luo et. *al.* is not connected to the status of KRAS.

2.2.4 Cell Lines used as Tumour Models

Another reason why the RNAi screening results for drug-target detection have not been successful in clinical trials, is the difference between the overall profile of cell lines, used in the in vitro experiments, and real tumours. A recent project funded by the Broad Institute, the Cancer Cell Line Encyclopedia (CCLE), has investigated the common characteristics between real tumours and approximately 1000 human cancer cell lines [49]. The comparison has been made in terms of shared mutations, DNA CNV, SNP e.t.c.. It has been found that the cell line mutation status, on which previous studies had been relying for comparison between cell lines and tumours, is not sufficient for proper transfer of the in vitro findings to clinics. A following study is inspired by the CCLE initiative and focuses in transfer of finidings in ovarian cancer from the lab to the clinics. It compares the commonly used for derivation of clinical conclusions ovarian cell lines, IGROV1, OC316, EFO27, OVK18 and TOV21G to high-grade serous ovarian cancer (HGSOC) tumours [50]. The comparison is conducted based on the Copy Number Alterations (CNA); the correlation of the CNA profile of each cell line to the tumours is calculated. It is astonishing that none of the most often used cell lines correlates well with the mean CNA tumour profile. This implies that there should be established new criteria for selection of cell lines on which experiments are performed.

2.3 Background in aggregation of ranks

One of the contributions of this thesis is consistent aggregation of ranked gene lists as calculated by various methods. The rank aggregation problem is concerned with the combination of many different orderings of the same elements, which are provided by different ranking schemes, in one list that best reflects the orderings of the underlying lists. The distance of the final list to the individual lists should be minimized.

This principle of rank aggregation is depicted in Figure 2.5.

This problem has troubled mathematicians in the past. Different approaches have been proposed, starting with Borda (1770), soon followed by Condorcet (1785). In the 1990's Arrow (1951) and Kemeny (1959) proposed small alternatives of the established methodologies, with the latter additionally satisfying the Pareto principle: In many events, about 80% of all the effects are due to 20% of the causes.

Two different philosophies have been formed throughout the years:

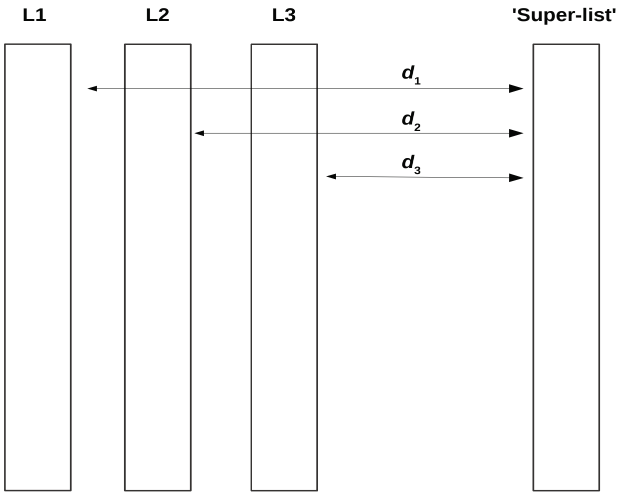


FIGURE 2.5: The idea is the detection of a super list that has as smaller distance as possible to the input lists. This figure presents and example of 3 lists, $L1, L2$ and $L3$ that are being aggregated in the "Super-list", d_1, d_2 and d_3 are the distances of each of the individual lists to the final one, and the rank aggregation procedure tries to minimize them

- Majoritarian Principle: This principle follows the "Condorcet" criterion: If $rank(A) > rank(B)$ more often than not in the input lists, then $rank(A) > rank(B)$ in the final list.
- Consensus among individual ordered lists: The rank of an item x in the final list is a consensus function of its ranks in the individual lists. Simple such consensus functions are the minimum, maximum, average, mean and others of the individual ranks of x. The Borda count, which is the average of the ranks of x in the individual lists, is a good representative of this category.

The selection of the best approach depends on the specific problem. Usually different approaches have different results.

The prevailing property of an aggregated ranking is the satisfaction of the "Condorcet's criterion". A natural step towards effective rank aggregation was given by Kemeny, who introduced the "Kemeny optimal ordering" [51]: Given k orderings $\tau_1, ..., \tau_k$, a *Kemeny optimal* ordering σ minimizes the sum of the bubble sort distances

$$\sum_{i=1}^{k} K(\sigma, \tau_i)$$

However, Kemeny optimal without further reduction is a NP-hard problem, even for the aggregation of 4 lists [51]. Dwork et al in [51] suggest a more relaxed version of the Kemeny optimal criteria; local Kemeny optimal aggregation, which still satisfies Condorcet's criteria and can be computed in O($knlogn$) time, where k is the number of input lists and n is the length of the lists.

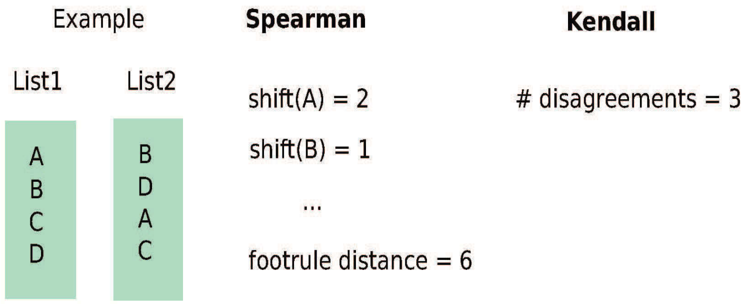


FIGURE 2.6: This small example shows two lists that contain the same elements in different order. The Spearman footrule distance counts what is the shift of each element between the lists. Kendall's tau just counts the number of disagreements of an items ordering between the input lists. So, for this example, in List 2 B is before A, D is before A and before C, though in List one their order is opposite. Thus there are 3 disagreements.

There are different types of distance measures that can be used to measure the distance between the individual lists and the *SL*. The most prevalent among them are Spearman's footrule (2.1) and Kendall's tau rank distances (2.2). The smaller the distance the bigger the similarity between the lists. Spearman's footrule measures the total element-wise displacement from the identity permutation Given a permutation σ on n elements, Spearman's footrule distance F(σ) is the sum of the absolute differences between i and $\sigma(i)$ over all values of i.

$$F(\sigma) = \sum_{i=1}^{n} |\sigma(i) - i| \tag{2.1}$$

The Kendall tau metric counts the number of pairwise inversions between two ranking lists.

$$K(\sigma) = \sum_{(i,j):i>j}^{n} [\sigma(i) < \sigma(j)] \tag{2.2}$$

Their similarity is that they both consider the ordering of lists. Their main difference is that Spearman additionally takes into account the amount of elements between the different ranks of item i in the lists. Moreover, Spearman is shown to be a good approximation for local Kemeny optimal aggregation. An example is provided in Figure 2.6.

Chapter 3

Data and Methods

In this Chapter, the datasets used in this thesis are initially presented. One of the most serious problems with gene ranking is the existence of many shRNAs per gene. Different techniques of shRNA-to-gene assignment are explored and the way in which each of them contributes to final gene ranking is shown. The overall methodology is divided in two parts. This Chapter is the first part of the analysis, and describes the methods that are applied in order to retrieve a first set of candidate KRAS SLPs. It also investigates the efects of the inclusion of an additional or a random screen in the analysis pipeline. The second part is essentially the evaluation of the findings of the first part and is covered in the Results Chapter. Since none of the techniques discussed is fully reliable by its own, the final results are the combination of all methods.

3.1 The datasets

The current analysis is based on three RNAi screens that were performed on isogenic colon cancer cell lines [25–27]. Colon (or *colorectal*) cancer, is the type of cancer with the second highest percentages of activating KRAS mutations after pancreatic, reaching 30 – 40% [52]. Initially, a larger number of screens had been collected for analysis by RankSLP. The final analysis was restricted to these three datasets as an effort to confront the high genomic variability of cell lines among different screens. It was impossible to account for all possible mutations, SNPs and CNVs of the cell lines involved in the screens. Each cell line's phenotype would be certainly dependent on all of these parameters and not only on the KRAS mutation. As RankSLP was interested in the effect of the KRAS mutation alone on the phenotype, it was decided to focus on clear datasets consisting of pairs of isogenic cell lines; they differ only in the existence or not of KRAS mutation.

TABLE 3.1: RNAi datasetes

Author	**Source**	**N° of targeted genes**	**N° of cell lines**
Luo *et.* al	Cell, 2009 [25]	19569	2 isogenic: $KRAS^{wt}: DLD1^{+/-}$ $KRAS^{mut}: DLD1^{-/-}$
Wang *et.* al	Oncogene, 2010 [27]	1740	2 isogenic: $KRAS^{wt}: HKE3$ $KRAS^{mut}: HCT116$
Steckel *et.* al	Cell Research, 2012 [26]	7283	2 isogenic: $KRAS^{wt}: HKE3$ $KRAS^{mut}: HCT116$

Table 3.1 summarizes the datasets on which the analysis was based. The genes claimed in this Table are calculated based on known - HGNC gene identifiers. Some more information about each screen is provided hereafter:

- **Luo screen**[25] : This is a genome-wide screen. The original data were provided after direct correspondence with the authors. They consist of fold changes, corresponding to viability values, for three biological replicates of each cell line. The data were already sample-wise normalized and log-transformed.
- **Wang screen**[27]: Raw data were provided by a third party[1]. This screen targets approximately 2000 genes, covering the majority of known human cancer genes and protein kinases. As in the Luo screen, the data consist of log-transformed fold changes, corresponding to viability values, for three biological replicates of each cell line, but they are not cell-line normalized. Thus, after initial data collection, they were subjected to sample-wise MAD normalization: For each sample, its median $median^s$ and its standard deviation sd^s were calculated. Then, for each sample s its i^{th} value v_i^s was transformed as:

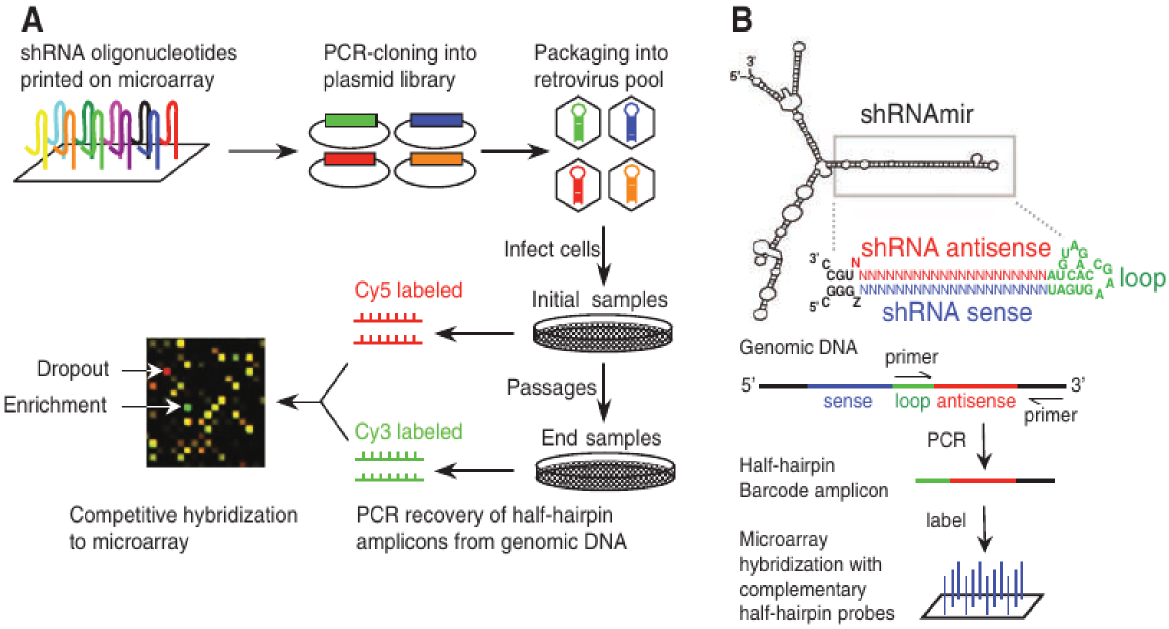
  $$v_i^s\prime = \frac{v_i^s - median^s}{sd^s} \tag{3.1}$$

  Moreover, plate-wise normalization was attempted for the Wang dataset. The results were very similar to the ones retrieved by cell line-wise normalization. In further computations, the latter was used.
- **Steckel screen**[26]: The data are freely provided as supplementary information to the publication and were downloaded from the web. Approximately 7000 human druggable genes are knocked down by siRNA pools. This screen also contains replicates but the provided data are also summarized per cell line (and normalized); thus each cell line (sample) value is the mean of its three replicates. The data are reported by the authors to be cell line-normalized.

---

[1] Alok Jaiswal, FIMM-EMBL PhD student, alok.jaiswal@helsinki.fi

The Luo and Wang screens are performed based on the protocol described by Schlabach *et* al. in [53], depicted schematically in Figure 3.1. Six shRNAs pools are used to silence the target genes. The same hairpin barcoded shRNAs are transfected into viruses which then infect with a 50:50 ratio both the KRAS mutant (GFP^+) and the KRAS wild type cells (GFP^-). The relative ratio of KRAS mutant versus KRAS wild type cells is examined at approximately seven days post infection. The shRNAs that are induced to the surviving cells are then rescued, whereas the shRNAs that are introduced in the cells that died are depleted from the population. Afterwards, the rescued barcodes are PCR amplified and are used to conduct a two-color microrray experiment. For the Luo screen, custom microarrays containing the HH barcode probe sequences, provided by Roche Nimblegen, are used. For the Wang screen, the microarray experiments are performed by Affymetrix Human Genome U133A 2.0 arrays using standard Affymetrix protocols. The final readout are the values from the microarray, bringing along background noise and printing errors, inherent in microarray technology. The genes corresponding to the rescued shRNAs are in abundance, and thus they are overexpressed on the array (Cy5 signal). On the other hand, the genes represented by the depleted shRNAs are underexpressed (Cy3 signal). Candidate KRAS SLPs are the genes of which the shRNA insertion has led to KRAS mut cell death after virus infection, thus the genes having a negative log fold change. This corresponds to green fluorescent microarray signal.


A
shRNA oligonucleotides printed on microarray
PCR-cloning into plasmid library
Packaging into retrovirus pool
Infect cells
Initial samples
Cy5 labeled
Passages
End samples
Cy3 labeled
Dropout
Enrichment
Competitive hybridization to microarray
PCR recovery of half-hairpin amplicons from genomic DNA
B
3'
5'
shRNAmir
shRNA antisense
loop
shRNA sense
CGU
GGG
AUCAC
UAGUG
Genomic DNA
primer
5'
3'
sense
loop
antisense
primer
PCR
Half-hairpin Barcode amplicon
label
Microarray hybridization with complementary half-hairpin probes

---

FIGURE 3.1: Figure and caption taken from [53]. Overview of the pool-based dropout screen with barcode microarrays. (a) Schematic of library construction and screening protocol, (b) Schematic of the HH barcode hybridization

The Steckel screen is performed in a different way: Pools of four siRNAs are used against each gene ("SMARTpools" from Dharmacon). siRNAs against KRAS are used as internal control. The readout of this assay is the fluorescence value of specific proteins which are induced during apoptosis, for example caspases. Fluorescence is read on an EnVision 2012 Plate-reader. This type of readout comes along with inherent mechanical errors accompanied by noisy readout signal.

3.2 Assignment of a single value in the case of multiple shRNAs per gene.

In the Luo and Wang screens, each gene is targeted by multiple numbers of shRNAs. This was treated with caution by the ranking algorithms that were used to prioritize the candidate genes. The assignment of only one value per gene was of most importance, in order to enable us to assess the gene's potency as a KRAS SLP candidate. This value should be representative of the real effect of the knock down of the intended gene. There are a number of algorithms fit for this shRNA-gene-value assignment. What should be taken into consideration when opting for which algorithm to apply is:

- The background distribution. Evaluation against the background distribution adds supporting evidence that the findings are different from random and accounts for the decrease in the number of False Positives (FP).
- The difference between the two classes, mutant and wild type. This is of great importance because the interesting genes are not just the ones of which the knock down leads the KRAS mutant cells to death. It should additionally be requested that the same genes don't lead the wild type cells to death. The desired effect (death) must be mutant KRAS specific.

In this thesis, the problem was approached by the application of three different methods, which took into account the above characteristics: A straightforward Standard method, the RSA and the RIGER methods.

3.2.1 Standard method

The Standard method simply ranks the genes based on how strongly their knock-down decreases the viability of the tumour cells, and selects the genes that score below a threshold. The search space was not yet limited at this first step in order to avoid early restriction of the dataset, as further steps and validations would follow. Therefore,

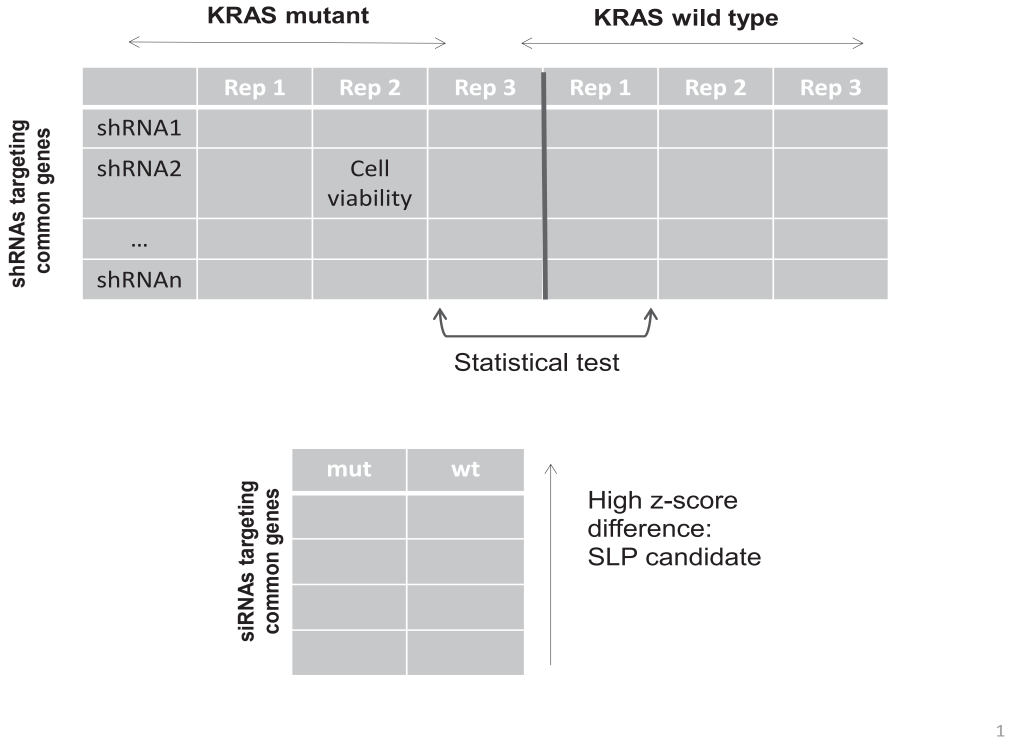


FIGURE 3.2: Top: Wang and Luo ranking approach. We apply the paired t-test between the columns. Bottom: Steckel dataset ranking. This is an apoptosis screen, so the higher the z-score difference (mut-wt) for a gene, the more probable KRAS SLP the gene is

all the shRNAs targeting a gene were included and ranked based on their differential effect on the viability of the KRAS mutant versus KRAS wild type cells. Of interest are the genes of which the depletion decreases the viability of the KRAS mutant cells but not the viability of the KRAS wild type cells. This can easily be captured by a statistical test which compares the means between the two conditions, in the case that a sufficient number of replicate experiments has been performed. As far as the *Luo* and *Wang* screens are concerned, they contain three replicate measurements for each cell line. Having only three replicates per cell line, normal distribution was safely assumed. The three mutant and three wild type replicates are connected, as they occupy the same well on the plate when the experiment is conducted. Hence, the *paired t-test* was applied between the two samples:

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}$$

In the above equation $\overline{X}_D$ and s_D are respectively the mean and standard deviation of the difference D between the two pairs of samples (mutant and wild type KRAS in this case). μ_0 is a constant used only for testing if the average of the differences is significantly different from μ_0. This was not important in the current case, so it was set to 0. Finally, n is the number of samples in each category.

The null hypothesis tested, was that depletion of a specific shRNA has no significant effect on the viability between KRAS mutant and KRAS wild type cell lines. The respective alternative hypothesis was that the viability of the KRAS mutant cells decreases more than the viability of their isogenic wild type counterparts. This is illustrated in figure 3.2. According to the Standard method, a gene was reported as hit when the best-ranking of all the shRNAs that target it was below threshold. The amount of shRNAs against the same gene, which rank below threshold, was further considered. The hypothesis that RankSLP made here, was that the more shRNAs targeting the same gene rank below threshold, the stronger the support that this gene is a true positive, i.e a more probable candidate for KRAS SLP (see section 4.6). Attention was paid to the trade off between the number of shRNAs targeting a gene and their frequency below the selected threshold.

As far as the Steckel screen is concerned, no replicate data are provided, as explained at the beginning of this Chapter. Thus, only the difference of the z-scores (mutant-wild type) was considered. This summarized value has the disadvantage that it doesn't take into account the variance variability, since it actually is a mean value. It has to be noticed that this is an apoptosis screen, so the result was inverted in order to be comparable to the other two viability screens. The rank R in this case was calculated as:

$$R = -rank(z.score_{mut} - z.score_{wt})$$

Calculation of the ranks by the Standard method was very quick ($O(nk)$), where n is the number of lists and k is the length of each one. This process provided a first estimation of the importance of each gene to the survival of a KRAS mutant cancer cell. Other methods followed, yelding supporting evidence for some of the highly ranking genes.

3.2.2 RIGER

The RIGER method was introduced in 2008 by Luo *et.* al. [54] and it is provided by the BROAD Institute of MIT and Harvard[2]. It is used towards the detection of hit genes in RNAi screens, in the case that each gene is targeted by multiple shRNAs. Thus, in the current data, it is applicable only for the Luo and Wang screens. RIGER calculates the ranking of a gene based on the averaged ranking of the two shRNAs with the strongest

---

[2] `http://www.broadinstitute.org/`)

differential effect between the viabilities of the mutants and the wild type cells. The respective algorithm has been incorporated into the freely available GENE-E software[3].

The steps that were followed by the RIGER algorithm are:

- Feature selection: each shRNA is scored according to its differential effect between two classes, here KRAS mutant and wild type. The RIGER tool provides a selection of methods for this step, like classical log-fold change criteria and t-test. To RankSLP's aim, the scoring was done based on the signal-to-noise metric as suggested by the authors. The data consist of two classes, mutant and wild type, having three samples in each category. This can be represented by a vector $c = (0, 0, 0, 1, 1, 1)$, which contains the labels for each sample, with 0 corresponding to mutant and 0 corresponding to wild type. The means μ_0, μ_1 and the standard deviations σ_0^2, σ_1^2 are the means and standard deviations of the mutant and wild type class respectively. For each shRNA, the signal to noise ratio is given by (3.2)

  $$SNR = \frac{\mu_0 - \mu_1}{\sigma_0^2 + \sigma_1^2} \tag{3.2}$$

  Only the genes selected at this step were forwarded to the second stage.

- Calculation of raw enrichment score: This is done in the same way as for the **G**ene **S**et **E**nrichment **A**nalysis (GSEA) [55], also provided by the Broad Institute. To this end, a gene-score assignment is performed by calculation of the weighted sum of the first two ranks of hairpins for a gene. Another approach that is provided by the RIGER tool is the weighted KS statistic (Kolmogorov-Snirnov). This statistic represents the degree to which these hairpins are overrepresented at the top or bottom of the ranked list of hairpins in the dataset. RankSLP was interested in the selection of genes that are overrepresented at the top of the list, that is why the weighted sum approach was chosen.

- Calculation of a normalized RIGER score: As a last step, the RIGER method normalizes the raw enrichment score (ES) to account for variable numbers of shRNAs across different genes. A background distribution was considered here, by 1000 random permutations of a hairpin set of the same size. Each gene had to be supported by at least two shRNAs for being subjected to normalization.

RIGER has the advantage of being a very quick procedure. The source code is not available but the real calculation time for RankSLP was between five and ten seconds.

---

[3] `http://www.broadinstitute.org/cancer/software/GENE-E/`

3.2.3 RSA

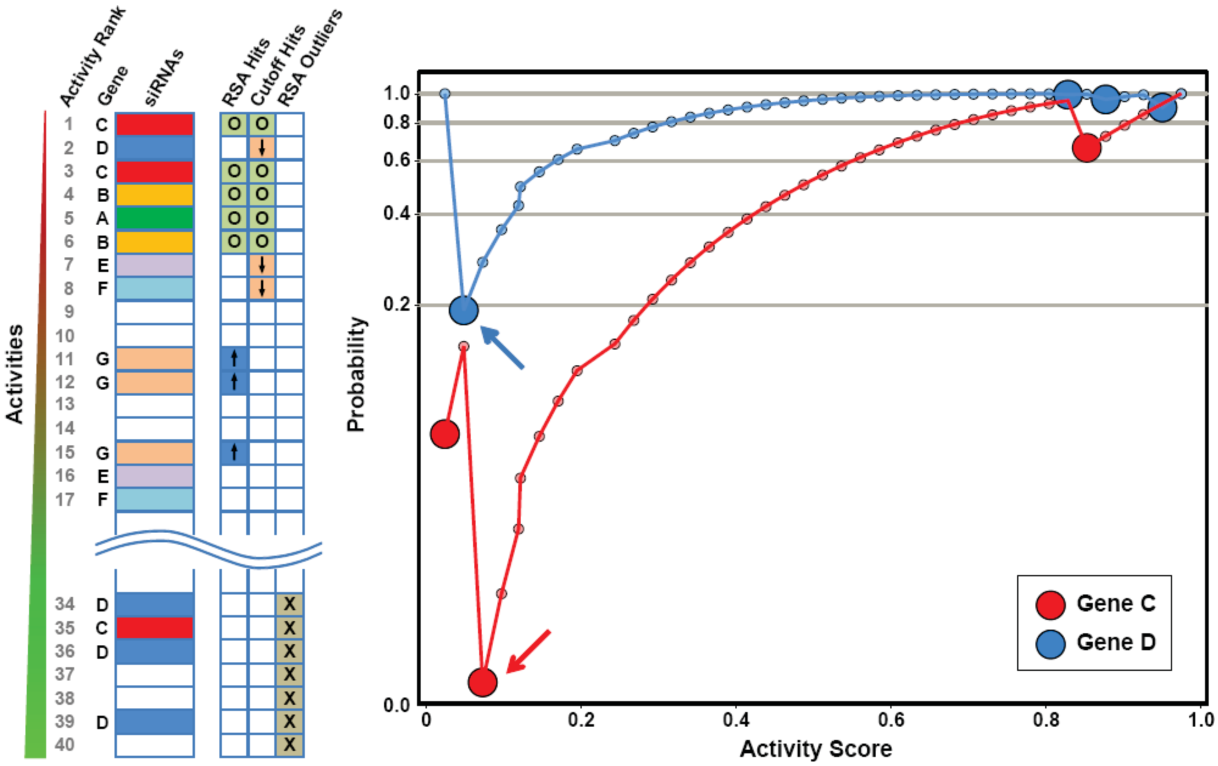


Figure 3.3: Illustration of RSA algorithm in siRNA hit selection. (a) Forty siRNAs are ranked according to their activities (potent on top) and colored according to their target gene identities. The top eight hits by both RSA and Cutoff algorithms are highlighted, with five common hits marked as "O", RSA-only hits as ↑ and Cutoff-only hits as ↓. siRNAs identified as outliers by RSA are marked as "X". (b) Iterative RSA p-value calculation process as illustration by Gene C (3 siRNAs) and Gene D (4 siRNAs). For a given gene, accumulative hypergeometric p-values are calculated for each siRNA, the curve dips at each siRNA targeting the gene itself (big filled circle). The global minimum is then identified (indicated by arrow) and separate siRNAs into two groups: hits and outliers. One and three least potent siRNAs are identified as outliers for Gene C and D, respectively. Gene C achieves a global minimum of 0.01, much lower than the 0.2 for Gene D, therefore, the activity distribution of Gene C is much less likely to occur by chance, therefore the gene is more likely to be confirmed. Figure and caption are taken from the RSA tutorial (`http://tex.stackexchange.com/questions/35043`)

RSA stands for **R**edundant **si**RNA **A**ctivity. The respective algorithm was introduced by Koenig *et.* al. in 2007 [56]. It is applied towards the detection of hit genes from RNAi screens in the cases where the same gene is targeted by multiple shRNAs or siRNAs. The final score of the gene is calculated based on the collective activities of all its shRNAs or siRNAs and a p-value is attributed to it. This p-value indicates how significant it is that all wells targeting the same gene are distributed among the higher ranking slots and it is computed by an iterative hypergeometric distribution formula. Intuitively, RSA assigns higher scores to genes that are targeted by more than one shRNAs with moderate activity, than to the genes with only one very active shRNA.

RSA belongs to the Gene Set Analysis (GSA) methods. It incorporates a "maximum mean" statistic for increasing the score of a gene, by calculating the absolute value of the mean of the gene scores in the mutant and in the wild type cells. The highest score is assigned to the category with the largest value. Efron and Tibshirani in their 2006 paper [57] show that this "maximum mean" statistic is often more powerful than the modified Kolmogorov-Smirnov statistic used in GSEA. RSA additionally calculates a null distribution by permuting both the genes (rows) and the samples (columns), which is different from sole column permutation (done in GSEA).

The R version of the algorithm was used. It was downloaded from the official website[4].The implemented algorithm iterates twice through the length of the input lists n and it computes the accumulated hypergeometric distribution function, which depends on the total number k of shRNAs in the library. Thus the final computation time was $O(nk)+O(n^2)$. Based on the same source, an illustrative example of the RSA algorithm is provided in Figure 3.3.

3.3 RNAiCut

RNAiCut is an algorithm, result of a collaborative project between Massachusetts Institute of Technology (MIT) and Harvard Medical School [58]. It is developed specifically for assessment of RNAi screen findings. A very common question when examining an RNAi screen's findings is *"Where to cut?"*. RNAiCut is aiming at the robust selection of a significance threshold for genes examined in such a screen, by combining it with the underlying PPI network information. The algorithm is based on the hypothesis that the true hits should be connected in an underlying network. Thus, for each set of top k genes, it calculates the p-value of finding a connected PPI subgraph of at least size k by chance. The lower the p-value, the more significant the input list k. This calculation was done quite fast by RankSLP (30 seconds for each list).

This resource was used as a complementary approach for threshold selection. It was applied on the ranked lists of the shRNAs which target the 1069 common genes of the screens Luo, Steckel and Wang. The Luo and Wang screens contain multiple shRNAs targeting these genes. None of them was excluded or pre-selected, for accordance with the previous steps of RankSLP. In this case of repeated genes in the input list, the algorithm internally discards each listing of the gene except for the first one.

---

[4] http://carrier.gnf.org/publications/RSA/

3.4 Rank Aggregation

In section 3.2.1, it is described how the rankings of each list were calculated using the Standard method. In this part, the interest is in finding a ranked list (super-list SL) that is as close as possible to each of the three individual rankings and best reflects the results from the ordered lists. This means that the final list minimizes the number of total disagreements. This procedure was applied only in the output rankings of the Standard method because of the need for broader applicability of RankSLP's findings; Standard is the only method which can be implemented on all three screens.

3.4.1 Sophisticated Rank Aggregation using heuristics

Rank aggregation is an optimization problem. One sophisticated way to generate candidate "super-lists" is by using heuristic functions. To this aim the RankAggreg R method was used [59], which incorporates two iterative heuristics:

- The **G**enetic **A**lgorithm (GA): This algorithm, mimics the natural selection process. As chromosome parts can be exchanged, and genes can be deleted, inserted or just replaced by others, in the same way, the elements of the input lists can be rearranged. This algorithm tests different rearrangements of the lists based on the input parameters. The most important are: 1) Cross over probability (CP): Two ordered lists can interchange their last slots which start at a random position with the CP probability, similar to how chromosomes interchange their tails. 2) Mutation rate (MR): Cross overs allow only for the re-arrangement of the lists. Mutations are more drastic events that will radically change the population, since they can completely alter the elements of the lists. Mutations can happen with probability MP. More details are provided in the book [60].
- The **C**ross **E**ntropy Monte Carlo algorithm (CE): Essentially the n input ordered lists can be re-arranged, and an element replaces another with a certain probability. This reminds a bit of a Markov Chain of states, where the states are replacing each other with different probabilities. As the Markov Chain develops, these probabilities get stabilized. That is the *stationary* property of a Markov Chain and the CE algorithm sorts the n candidate lists based on these probabilities. At each stage it generates a random data sample from the input lists. It then updates the parameters of the model based on the data. The aim is to produce a 'better' sample in the next iteration by reducing cross-entropy. More details on the algorithm are provided in the respective paper [61].

Both heuristics are iterative. The methods converge when the optimal "super-list" remains optimal for a number of consecutive iterations. They both require a distance measure which calculates the similarity of each iteration's output list with each of the input lists. Spearman's footrule distance was chosen because i) it is the one which takes into consideration the distance between the ranks of the same element across lists, as opposed to the Kendall's tau metric and ii) it is a very good approximation of Kemeny's local optimization (NP-hard problem), but can instead be computed in polynomial time ($O(n^3)$), where n is the amount of element in the lists. RankSLP's main interest was essentially the retrieval of a robust set of genes at the top of the super-list. It thus aggregated the top 10% hits of the three screens which translates to 107 genes (initial number of common genes is 1069). It has to be noticed that the original paper [59] clearly points out that the *RankAggreg()* function does not guarantee an optimal solution for a large number of items (where large is defined as more than around 100). This is a supplementary reason why the ranking of only the top 107 genes were aggregated.

Once the genes from the aggregation of the top 10% of the lists were retrieved, RankSLP examined which of them are even more robust by aggregating the top 5% of the ranked lists.

Last but not least, this package encourages the use of importance weights on the input lists. This is reasonable; not all of them should be equally taken into account if it is suspected that one is of better quality than the others. The selection of a representative importance scheme is examined in section 4.8

Details on the complexity time of the algorithm are not provided but generally, the heuristics that make use of Markov Chain property are $O(n^3)$, where n is the amount of elements in each list. Considering that the calculation of the Spearman Distance is of polynomial time $O(n)$, the total algorithm is $O(n^3)$. The real time that is consumed in this calculation is about six hours in one Core of an Intel X5650 CPU, 2.83Hz, 96GB of memory.

3.5 Random or additional screens: How much do they alter the result?

A novel methodology which explores whether the incorporation of an extra or a random screen alters the retrieved set of significant genes was developed. The aim is to judje the robustness of the approach followed in this thesis. To this end, a table with the ranks of each gene in each screen was created. An example case is visualized in Table 3.2. All the genes that appear at least once in the screens were considered.

Each screen's gene rank was assigned as described in Table 3.2 and further transformed to a percentage value by dividing it with the number of genes in the screen, excluding NAs. Thus, in the end, the table was filled with float values, which from now on we will call "relative ranks". Then the *Fisher* method for combining relative ranks was applied, using the *fisher.method* function from the *R MADAM* package. The approach as described by Fisher combines relative ranks to a statistic

$$S = -2 \sum^{k} logp,$$

which follows a χ^2 distribution with $2k$ degrees of freedom, where k is the number of tests being combined. All possible screen combinations were considered; one screen only, pairwise, three-wise and all four. The complexity of this process is $O(nk^2)$, where n is the length of the input lists and k is the amount of lists. The combined relative ranks were calculated and used in the proceeding steps.

Table 3.2: Each column corresponds to the ranks in the respective screen. The total number of genes (and consequently the rows of the table) is n, equal to the number of genes screened in Luo, because Luo is the genome-wide screen. Each gene may or may not be present in the rest of the screens. In case it is present, a value v_i is assigned to it, which is the rank of its best ranking shRNA, based on the Standard Method ranking. The best ranking shRNA was used for consistency with the general procedure that was followed in all the Standard and Rank Aggregation rankings. In case a gene is not present in one of the screens, the respective field takes the NA value. The fisher method, takes into account the NAs in the calculation of the summarized relative rank

Gene	**Luo**	**Wang**	**Steckel**	**Barbie/Random**
G_1	$v_{1,1}$	$v_{1,2}$	$v_{1,3}$	$v_{1,4}$
G_2	$v_{1,2}$	NA	$v_{1,3}$	NA
G_3	$v_{1,3}$	NA	NA	NA
G_m	...	...	...	...
G_n	$v_{1,n}$	NA	$v_{1,3}$	$v_{1,4}$

3.5.1 Extra screen: Barbie

Barbie *et* al. provided this screen in their publication [24]. It is maintained by the Broad Institute as *"Achilles 1"* dataset, because it is supposed to retrieve the Achilles heel set of genes which, if targeted, can kill a KRAS mutant cancer cell. 957 genes are targeted in 19 human cancer cell lines. Mutation status of KRAS is either complementary to the dataset or retrieved by COSMIC (see subsection 4.11.2). The obtained values are plate-wise normalized and provided as B-scores, which were retrieved from CanSar[5] after direct correspondence with Joe Tym, the CanSar database curator, on 28 March 2012.

---

[5] `https://cansar.icr.ac.uk/`

The cells were separated into KRAS mutant and KRAS wild type, as done with the rest of datasets (Figure 3.2). The cells are not isogenic in this case, but the existence or not of the KRAS mutation only was considered, being aware that this is a simplified case. The wilcoxon rank sum test was applied on each shRNA between the KRAS mutant and the KRAS wild type cells. The respective rank for each gene in Table 3.2 is the rank of its best ranking shRNA, for consistency with the Standard method.

3.5.2 Random screen

For the calculation of the random values, the original relative ranks for the three screens were pooled together. Afterwards, as many elements as the total amount of genes (equal to the size of the Luo screen) were sampled with replacement from this pool of values. This was regarded as our *random screen*.

Chapter 4

Results

At the beginning of this Chapter, more detailed information is provided on the existing screens that have detected potential KRAS synthetic lethal partners. The genes retrieved from each study are presented. Their small overlap was a strong motivation for the current work. Additionally, an overview of the data used and the integration problems that they expose is given, and ways in which the integration problems were encountered are described.

The core part of the results starts with exposing the results of the application of the Standard method, RIGER and RSA. These are followed by the findings of the three screens' rank aggregation and by RNAiCut. Among methods, some different and some common genes are retrieved. At the end, some more probable candidate KRAS SLPs are identified. The basis of the hit detection method used in this thesis, is the combination of different approaches to a final result. The more frequently a gene is retrieved when all methods are considered, the more potent it is.

The set of candidate concluding genes is further evaluated against the genes that are already reported in the literature as KRAS SLPs and against independent compound inhibition screens. Moreover, the modules of the enriched genes are represented on networks and interesting connections are further discussed. Finally, the candidate list is further limited based on the evaluation outcome. In total, the results of nine methods are reported and considered for the final gene selection: Standard method, RIGER, RSA, RankAggreg, RNAiCut, hypergeometric test on the significance of the genes of which two or three shRNAs rank above the threshold, evidence from literature, network enrichment and results from external drug screens. The retrieved genes' immunity to noise was also examined. Few promising genes are subject to experimental testing by our collaborators at the Medical Faculty.

The followed methodology in this work is high dimensional and involves the selection of many parameters. Each of the following sections explain the parameters chosen by each method and its results. Here, the whole procedure is summarize with the help of two figures 4.1 and 4.2.

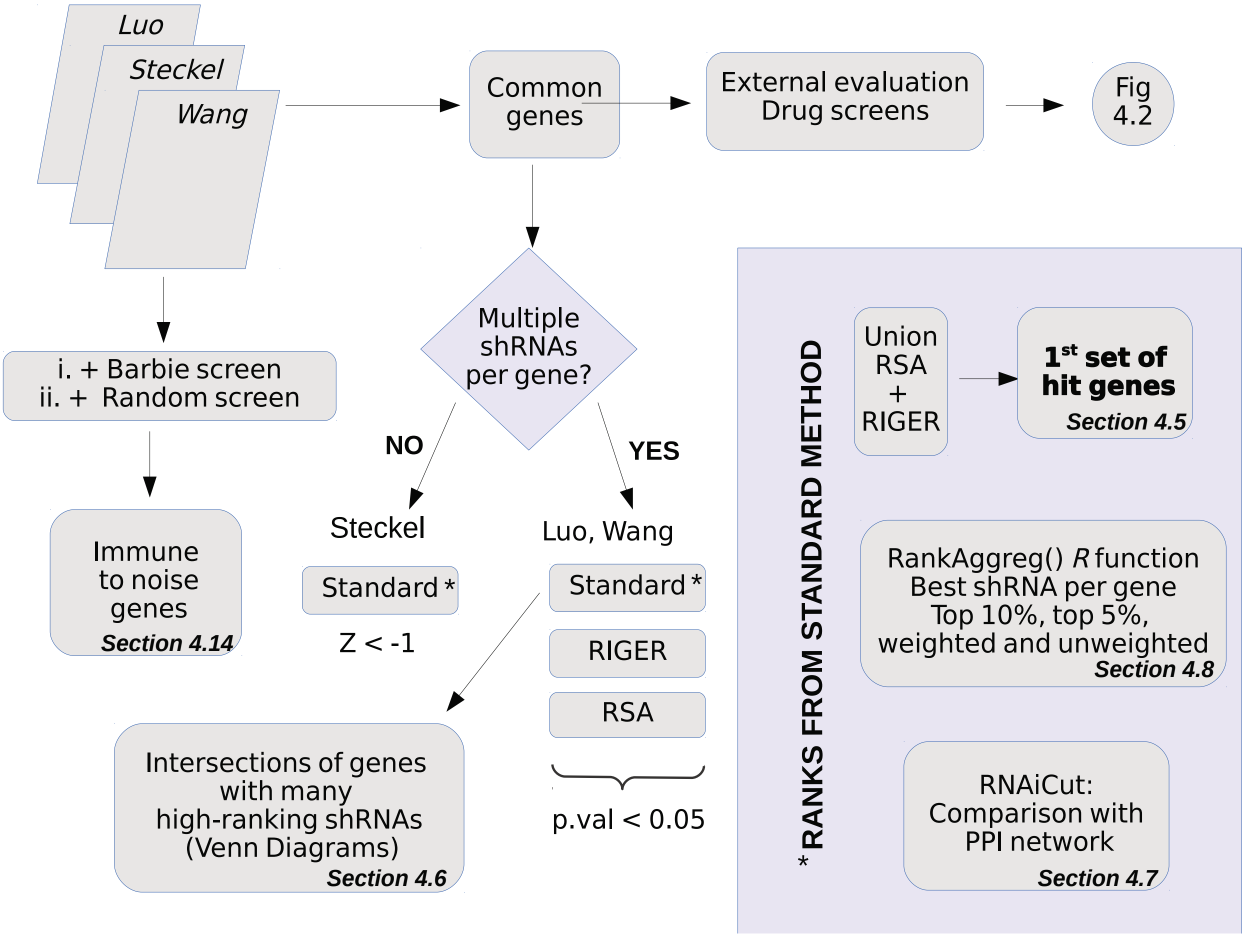


FIGURE 4.1: Approach: Schematic 1. The first set of methods that were followed for the analysis of the screens along with selected thresholds.

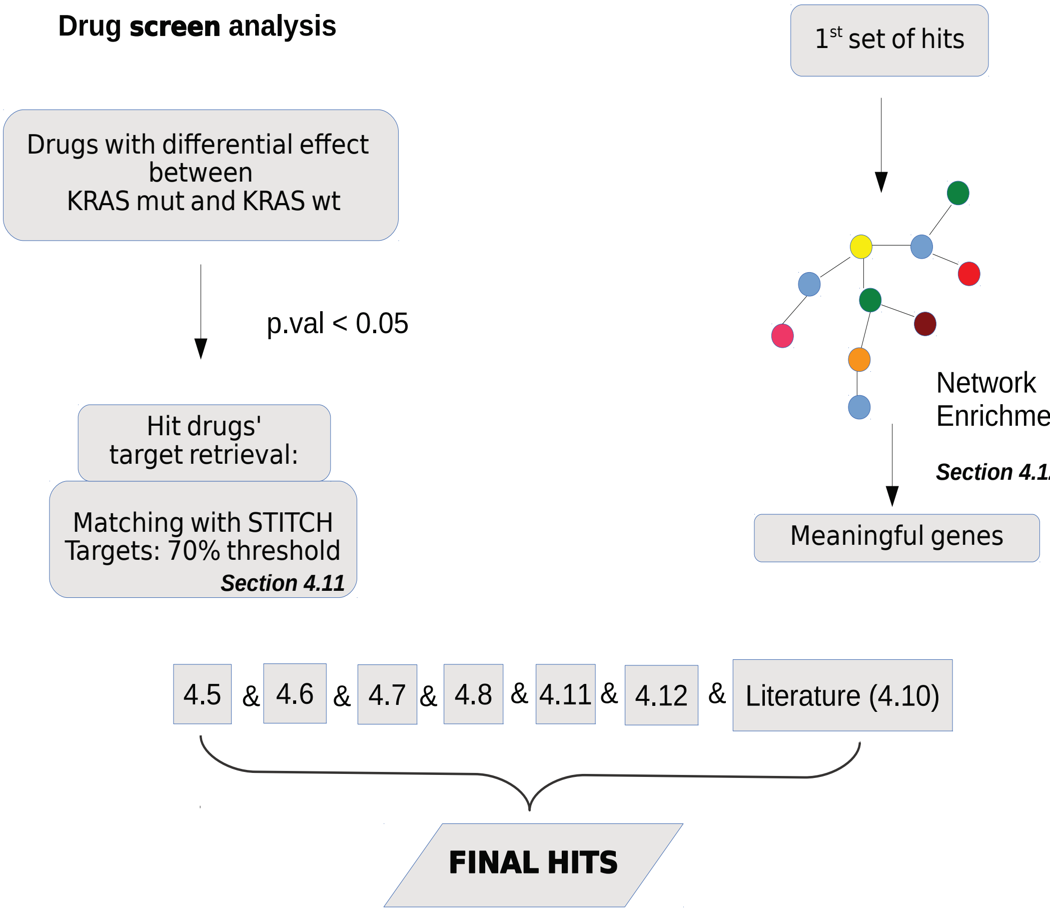

Drug screen analysis
Drugs with differential effect between KRAS mut and KRAS wt
p.val < 0.05
Hit drugs' target retrieval:
Matching with STITCH Targets: 70% threshold
Section 4.11
1st set of hits
Network Enrichment
Section 4.12
Meaningful genes
4.5 & 4.6 & 4.7 & 4.8 & 4.11 & 4.12 & Literature (4.10)
FINAL HITS

Figure 4.2: Approach: Schematic 2. The evaluation methods and the final gene selection

4.1 KRAS SLPs in Literature

As already stated, there are various experimental studies aiming at the identification of KRAS SLPs. Some of them are performed using high-throughput RNAi screening at first (genome-wide or more focused), followed by low-throughput experiments on the initial hits, for example Scholl *et.* al. [28]. Some others, are based on previous indications and focus just on low throughput experiments on specific cells [62]. In order to retrieve the KRAS SLPs we searched PubMed, initially with the query *"KRAS synthetic lethal"*. The result of the query, as of 17-03-2014, contained 28 papers, including KRAS SLPs in species different from human. Some of the retrieved studies detected drug d_i and gene g_i combinations as KRAS SLPs [63]. This means that a gene g_i is SLP of KRAS only under the presence of a specific drug d_i. To avoid missing any potential SLPs, the query was expanded to *"RAS synthetic lethal"*. This resulted in 68 papers, including KRAS and other members of the Ras-family (HRAS, NRAS) SLPs as well.

Table 4.1: KRAS SLPs in literature

KRAS SLPs	**From datasets** [1]	**From rest literature**
Alt-NHEJ pathway [64]		√
APC / C complex (e.g ANAPC1, ANAPC4, CDC16, CDC27) [1, 25, 26]	√	√
ATR [65]		√
BCL2 [66]		√
BCL2L1 [24]		√
BCL2L1 + MEKi [63]		√
BIRC5 [25, 46]		√
BRIX1 [25]	√	
cAMP / PKA pathway [67]		√
CASC5 [25]	√	
CCNA2 [25]	√	
CDC6 [26]	√	
CDCA8 [25]	√	
CDK4 [62]		√
CHEK1 [65, 68]		√
COPS3 [25]	√	
COPS4 [25]	√	
COPS8 [25]	√	
REL (encodes c-REL tr.factor) [24]		√

---

[1] The datasets that were incorporated in RankSLP

CUX1 [69]		√
DHX8 [25]	√	
SMAC + TRAIL2 [15, 70]		√
EIF3C [25]	√	
EIF3G [25]	√	
FBL [25]	√	
FIP1L1 [25]	√	
GATA2 [26, 71, 72]		√[2]
GSPT1 [25]	√	
HNRNPC [25]	√	
IL8 [73]		√
JAK1 [25]	√	
KIF2C [25]	√	
LDHA (only under hypoxia conditions) [74]		√
MAP3K7 [75]		√
METAP1 [25]	√	
MIS18A [25]	√	
NAE1 [25]	√	
NEDD8 [25]	√	
NFKB pathway [24]		√
NOL56 [25]	√	
NXF1 [25]	√	
OIP5 [25]	√	
PI3K-AKT-mTOR pathway [76, 77]		√
PKCδ (PRKCD) [78]		√
PLK1 [25]	√	
PSMA5 [25, 26]	√	
PSMB5 [25, 26]	√	
PSMB6 [25, 26]	√	
PSMD14 [26]	√	
RALB [24]		√
SAE1 [25]	√	
SIAH2 [73]		√
SMAD1 [75]		√
SMC4 [25]	√	
SNAI2 [27]	√	
STK33 [28]		√

---

[2] Although GATA2 comes form a publication from which some data were used in this thesis, GATA2 was evaluated on an independent dataset to which we worked with.

SYK [6]		✓
TBK1 [24]		✓
THOC1 [25]	✓	
TOP1 [26]	✓	
TOP2A [26]	✓	
TPX2 [25]	✓	
TWIST1 [79]		✓
UBA1 [25]	✓	
UBA2 [25]	✓	
UBE2I [25]	✓	
USP39 [25]	✓	
VDAC3 [80]		✓
WT1 (Hugo symbol: PAWR) [81]		✓

Table 4.1 contains published KRAS SLPs in alphabetic order, along with the respective publication. Many of them consist of secondary screen findings as there is just a hint in the respective publication that they may be KRAS SLPs. Subsections 4.1.1 and 4.1.2 highlight the KRAS SLPs from the RNAi screens that were used by RankSLP and from the rest of literature respectively.

4.1.1 KRAS SLPs from RNAi screens

Three high-throughput RNAi screens with available data are used for the current analysis by RankSLP. These studies are conventionally named as *Luo*, *Wang* and *Steckel*, based on the name of their first author. The datasets are extensively presented in section 4.3. The main KRAS SLP candidates that the authors of the respective papers identified were extracted from the second column of Table 4.1 (4 in total): PLK1 [25], SNAIL2 [27], CDC6 and GATA2 [26]

4.1.2 KRAS SLPs from the rest of literature

Studies without available datasets are kept as well, with the intention to be used as an independent set for the evaluation of RankSLP's findings. These studies come up with 24 different *modules* which are identified as KRAS SLPs and are indicated in the third column of Table 4.1. They are called *“modules”* because some of them are whole pathways. The respective genes are 28 in total: AKT1, AKT2, AKT3, APC/C, ATR, BCL2, BCL2L1, BIRC5, CDK4, CHEK1, CUX1, FRAP1, GATA2, IL8, LDHA, MAP3K7, NFKB1, PAWR, PRKCD, RALB, REL, SIAH2, SMAD1, STK33, SYK, TBK1, TWIST1, VDAC3. In the following Chapters, this set of genes will be referred to as *Gold Standard Genes* (GSGs). In this set of GSGs we intentionally don't include the 4 genes mentioned in 4.1.1, because they are coming from the screens that we analyze; thus they are not appropriate for use in validation of the developed methods.

4.1.3 Overlap

The next step was to examine if the existing screen findings agree. A comparison of the KRAS SLPs from literature was conducted, including both the SLPs in the datasets which are used by RankSLP and the ones without data. The observed agreement was partial: Only BIRC5 clearly agrees between only two studies. Components of the proteasome (APC/C, PSMA5, PSMB5, PSMB6) are also detected as KRAS SLPs by both Luo et. al [25] and Steckel et. al [26]. The novel idea that motivated this work is that there should be more SLPs in the screens that were underestimated at first place. RankSLP's goal is to apply integrative analysis approaches for retrieving genes that score consistently high in all the screens. The idea is that, even if they don't score at first positions in all of the screens, as long as they score relatively high this is undoubtedly stronger evidence that they are real SLPs. It is more probable that targeting one of them in a new KRAS mutant sample, will successfully drive it to lethality.

4.2 Getting to grips with the datasets

The datasets that were used in this integrative analysis are thoroughly presented at 3.1. A summary of the differences they expose is:

- Luo and Wang screens use shRNAs to silence the genes of interest, whereas the Steckel screen uses pools of four siRNAs.
- In Luo and Wang screens many shRNAs are used against the same gene. On the other hand, in the Steckel screen, the siRNA pools are not deconvoluted to their individual oligonucleotide molecules, thus only one value per gene is available.
- The Luo and Wang are viability screens; the viability of the cells, with respect to a specific shRNA knock down is measured. The Steckel screen is an apoptosis screen, meaning that the apoptosis of cells is measured
- Luo and Wang datasets essentially contain microarray values, corresponding to log-fold changes of the cases versus the controls. The Steckel dataset contains one value which is the normalized fluorescent signal of proteins that are released during apoptosis (i.e. caspases)

As explained in 3.1, the values of Luo and Steckel screens are already log-transformed and normalized. The Wang screen data were subjected to cell line-wise MAD normalization. Figure 4.3 depicts the distribution of the data in the final format in which they were used in proceeding analysis.

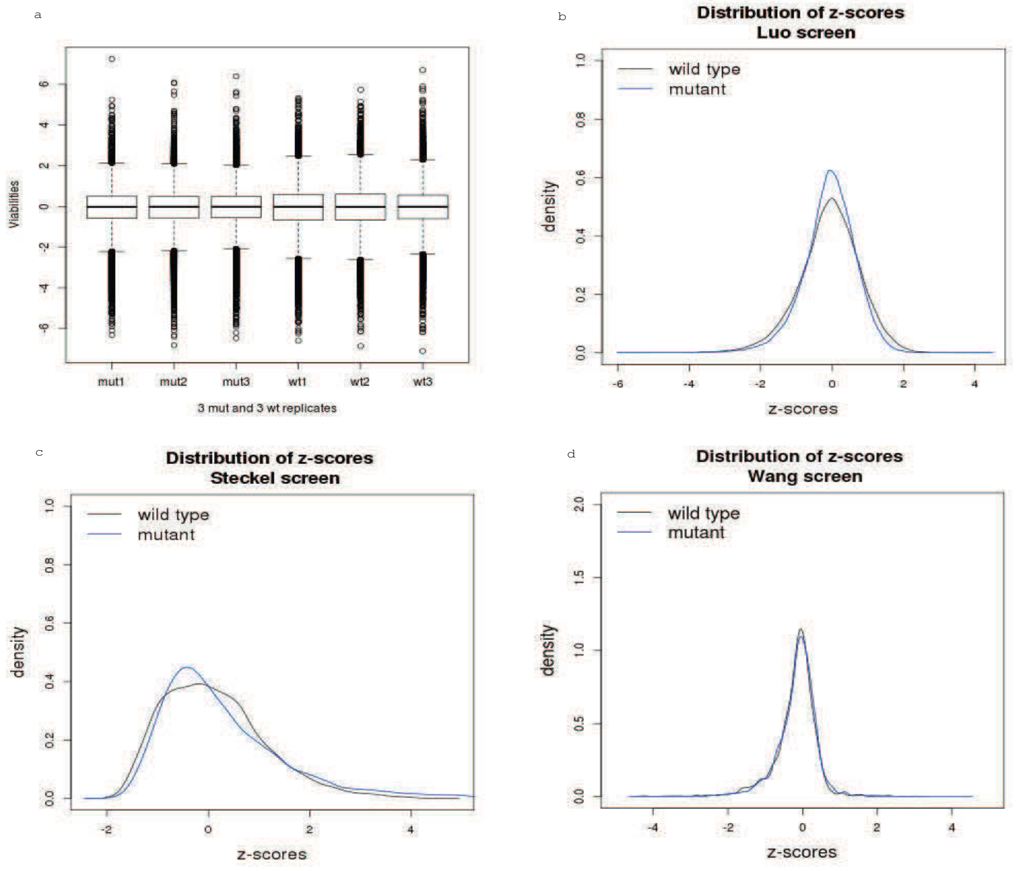


FIGURE 4.3: a. Cell viabilities for mutant and wild-type cell lines in the Luo screen. There are three wild type and three mutant replicates. b,c,d: The density of the normalized data for Luo, Steckel and Wang datasets respectively. The plots show the averaged values over the three replicates of each cell line.

4.2.1 Multiple shRNAs per gene

In Luo screen, the average number is five shRNAs per gene, however there are considerable differences among the genes. For example, there are genes (e.g. TBK1) that are targeted by two shRNAs, and genes that are targeted by seven shRNAs (e.g. COPS2). The problem is that the cell viability values corresponding to each shRNA against the same gene don't agree. Different shRNAs have differences in their effectiveness on knocking down the target gene, which happens mainly due to OTEs. To visualize one of the cases, the values of two of the three shRNAs targeting the same gene in the mutant cell lines of the Wang screen are plotted (Figure 4.4). To connect this with the actual microarray values and fluorescence, the fluorescence of three example genes in the Wang screen are also presented in Figure 4.5.

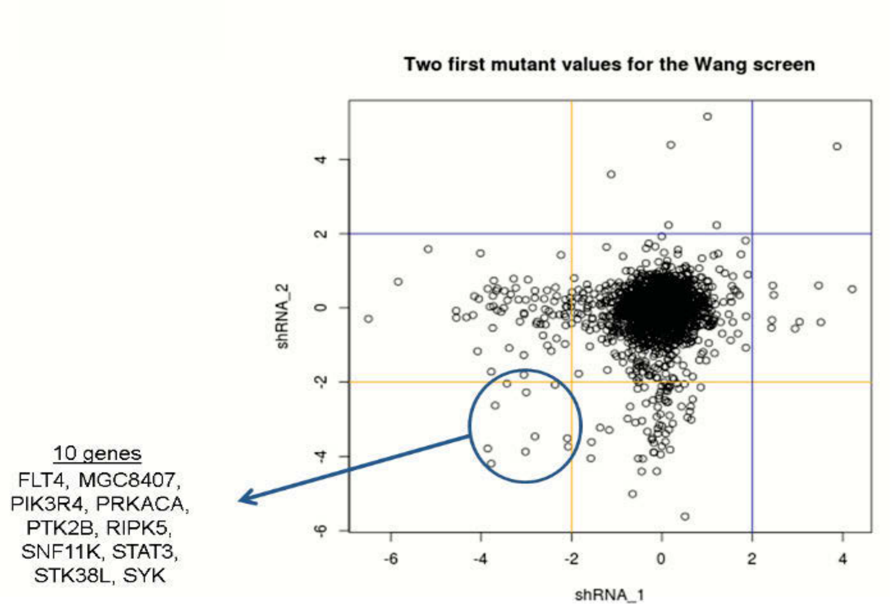


FIGURE 4.4: The values of the first and second shRNA (out of three in total) targeting a gene in the mutant cell lines in the Wang screen are plotted respectively on the x and y axis. Many shRNAs agree in values close to zero, as shown by the big bulk at the center of the plot, but these cases are uninformative to our analysis. There exist differences between the first two shRNAs for the rest of the values, except around only 10 out of the common 1070 genes. These are targeted by shRNAs of which the fold changes both agree to < -2 (depleted ones).

4.2.2 How differences in knock-down technologies and readout methods were encountered?

The above points raised the question: *"How to compare the different screens?"* Despite the variability among the screens, they essentially measure the same thing: How much a gene's knock-down influences the viability of the cancer cells and what is the observed difference between KRAS wild type and KRAS mutant cells. The numerical comparison of the screens was accomplished by normalizing all measurements (conversion to z-scores) and calculating based on them. The normalizations were performed cell line-wise as explained above. Gene-wise normalization was not needed because the data were not combined into a bigger dataset; only the independent gene ranks from each individual screen were needed.

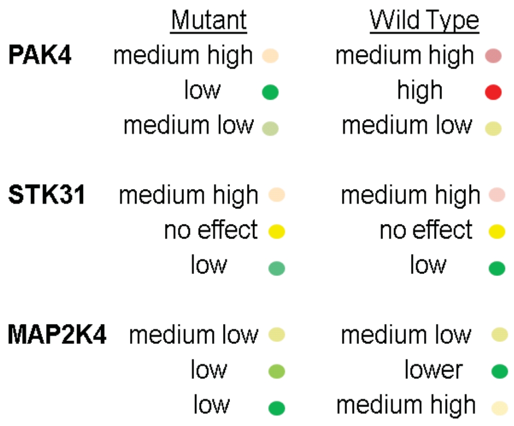


FIGURE 4.5: Fluorescence of three example genes, PAK4, STK31 and MAP2K4, on the microarray that measures the viability of the cell lines of the Wang dataset. The colors correspond to the Cy3(green) and Cy5(red) dyes that are used on the microarray, with green and red reflecting gene depletion and abundance respectively.

4.3 How was screen size variance encountered?

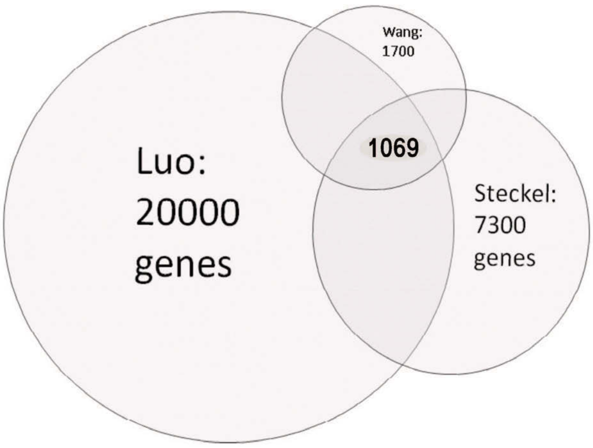


FIGURE 4.6: Overlap of the targeted genes in the three screens. Only 1069 genes are screened in all of them and these are the ones that were considered for further analysis in the current work. This means that ∼ 19000 genes were dropped.

The three screens that were used as basis for our analysis vary in size (Figure 4.6). Luo is genome-wide but the other two target only a part of the human genome. This variability in screen size was the first obstacle to RankSLP. Which genes should be considered? In order to retrieve results that are applicable in wide scale, which is the aim of this thesis, as many genes as possible need to be covered. However, comparison of the findings between the screens required focusing on the common genes only. We had of course many considerations before proceeding to this step. For example, if a gene is very promising in one screen but is not even examined in the others, RankSLP would miss it. This is a limitation of the current approach which is based on the hypothesis that a gene has to be important in all screens in order to be a KRAS SLP candidate. Thus, this type of genes was missed, but at least widest possible application of RankSLP's fidings was ensured.

The number of common genes among the three screens is 1069. To map the provided gene symbols to HGNC identifiers, the PubMed gene file[3] as of 25 February 2014 was used. The majority of them, 1026, were already HGNC identifiers. The rest 43 genes had the same identifier among the screens but this was not HGNC identifier. For screen overlap, the name appearing in the screens was used. Only two of the non-HGNC identifiers were hits by RankSLP's analysis: *MLL3* and *MDS1*. The respective HGNC identifiers are *KMT2C* and *MECOM*. The HGNC symbols of these two genes were used in further steps; annotation, enrichment and evaluation of hits using external data.

The next question was: Why three screens and not three or four, e.t.c.? The amounts of common genes between each screen pair and among all three screens are:
Luo - Steckel: 5807
Luo - Wang: 1423
Steckel - Wang: 1207
Luo - Steckel - Wang: 1069
As it can be observed, for two screens there is still a relatively high overlap. For three screens this drops very quickly and for four screens it would drop even more. Initially, an additional screen (Barbie) had been included, and the total overlap was 400 genes. Keeping the Barbie screen would have shrunk the searching space a lot. Then, why not just two screens? Because, as already explained, the main aim was the widest possible applicability of this work's findings. For two screens, the pairwise overlap is high but, two screens is a borderline number. When a third screen is included, the overlap is above 1000 genes, still not bad. Thus, three screens was considered a good number for counter-balance between general applicability and sample size.

4.4 How were cell line differences encountered?

The cell lines used in the selected experiments belong all to colorectal (colon) cancer. As described in table 3.1, the Steckel and Wang screens are performed on the same cell lines, *HCT116* (KRAS mutant) and *HKE3* (KRAS wild type). On the contrary, Luo screen is applied on *DLD1*: KRAS wild type and mutant. In the latter case, the mutation G13D at codon 13 of the KRAS gene is technically induced.

---

[3] `ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/`

Although the cell lines differ among the screens, they are comparable. The reason for that is that they all are isogenic, meaning that the respective mutant and wild type cells are similar, apart only from the KRAS mutation. In addition, all the mutants harbor the same KRAS G13D mutation in codon 13, where the amino acid *glycine* is substituted by *aspartic acid*. According to [82], G13D mutations are met in 19% of human colon cancers. Having isogenic cell lines is very advantageous for detecting differential genes because the only thing that differs is this mutation [83].

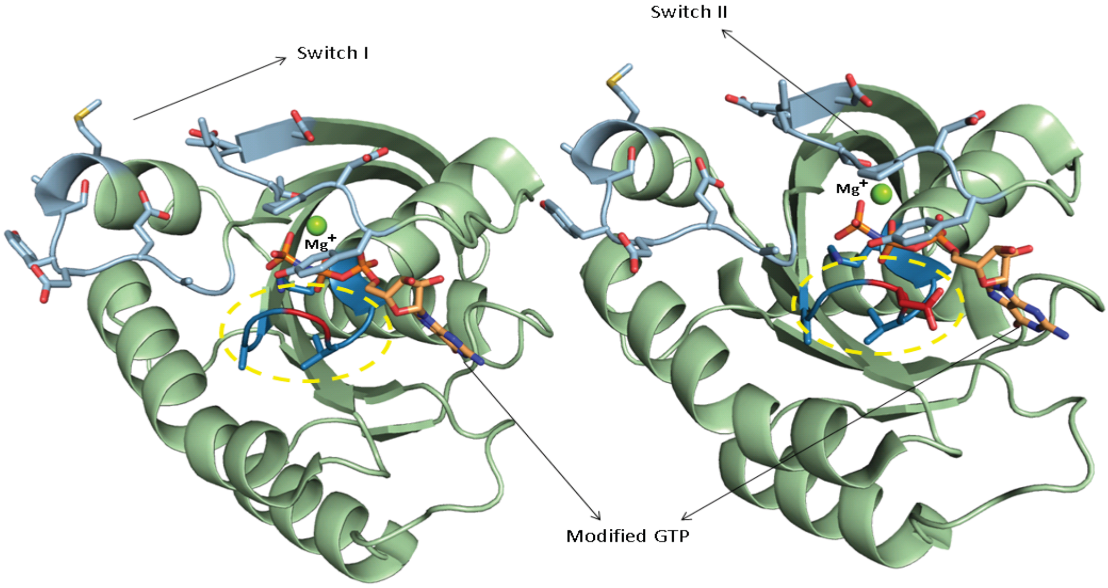


Figure 4.7: Structure of wt and mut KRAS, with codon 13 colored at GTP binding site. This figure is created using PyMOL[4] and shows the structures of KRAS G13D and KRAS wild type, with codon 13 colored at GTP binding site. The published crystal structure 3GFT from PDB was used. Left: wild type KRAS. Right: mutant G13D KRAS. The GDP /GTP binding site is circled by yellow dashed line. The bound GTP is shown in orange in both figures. The mutation happens at the P-loop of the KRAS molecule and leads to a dramatic change at the GTP-binding socket. This results in KRAS being constitutively locked to GTP.

4.5 Retrieved genes from the three main methods: Standard, RIGER and RSA

Technical details on the three main methods incorporated in RankSLP are provided in section 3.2. For the standard method and for the Luo and Wang datasets, all the shRNAs that target each gene were considered. RankSLP was interested in significant shRNAs, having a p-value < 0.05. Thus, this threshold criterion is applied on the shRNAs of the whole dataset.

For the Steckel screen, where only z-scores and no replicates are available, the standard criteria z-score *smaller* than a threshold was applied. First, the extensively used threshold -2 was selected. Due to small hit retrieval with this threshold, RankSLP also experimented with -1, which is the one that was finally used.

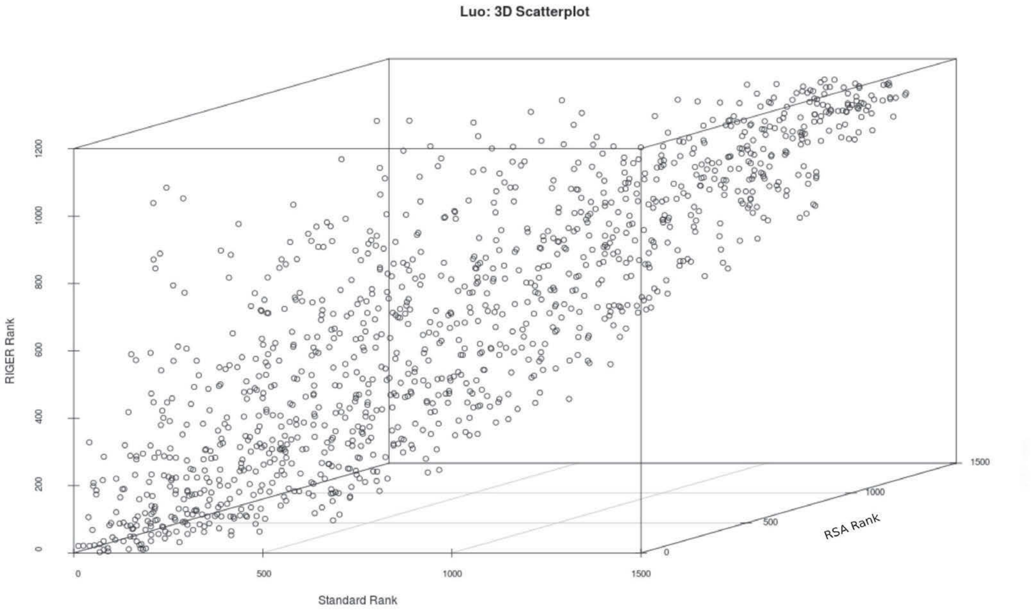
RIGER and RSA methods also return a p-value for each gene, after evaluating the contribution of the different shRNAs that target it. The RIGER algorithm calculates a p-value for each gene, indicative of the significance of the differential KRAS mutant and KRAS wild type cell viability, caused by the gene's knock down. To this aim a weighted sum approach based on the signal to noise ratio was utilized. For the RIGER method, a simple criteria of p-value < 0.05 was sufficient.

RSA method on the other hand assigns a value to each experimental well and in this case the selection criteria are a bit more complicated. The inputs provided to RSA are the summarized mutant-wild type values per well for Luo and for Wang screens. According to RSA, a lower bound LB and an upper bound UB are defined, between which hit genes are sought. These bounds intuitively correspond to fold changes. Wells with lower scores than LB are guaranteed hits, whereas wells with larger scores than UB are guaranteed non-hits. After correspondence with the methods' developers and matching of the LB and UB bounds with fold change, RankSLP opted for $LB = -2$ and $UB = 0$. There was increased confidence in favor of their being hits if the log transformed and normalized fold change is < -2. The ones having a positive fold change (> 0) correspond to the surviving cells, thus they were rejected. The most significant hits were selected based on two criteria:

- The p-value for each gene (p-values are the same for all wells corresponding to a gene) should be < 0.05, and
- The gene should have at least two active wells ($OPI\_Rank < 99999$ for at least two wells)

It has to be noticed that the rank of at least two wells, having p-value < 0.05, should be a real number (RSA returns infinite if a rank cannot be calculated).

The retrieved ranks were mapped on 3D axes, with each axis representing the rankings as calculated by each method (Figure 4.8). An obvious observation was that although the exact ranking number differed, there was a general trend of correlation. The pairwise Spearman correlation values were calculated for the rankings of the Luo and Wang screens. After this step, RankSLP proceeded to hit gene set selection by each of the three main methods, by applying a two-step procedure:



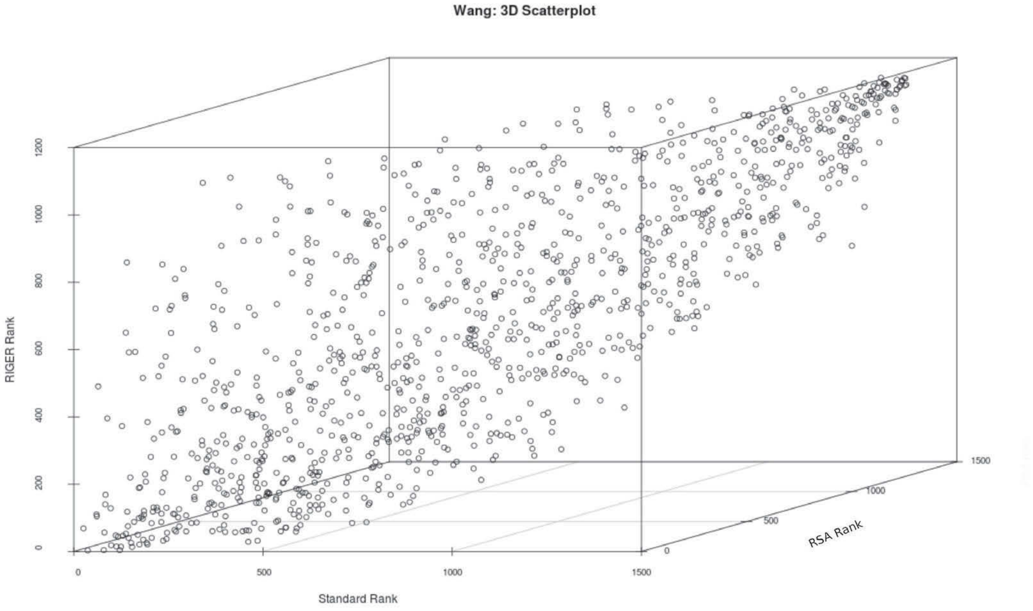


FIGURE 4.8: Scatterplots of the gene ranks retrieved by the three main methods for the Luo (top) and Wang (bottom) screens. The Standard, RIGER and RSA ranks are shown on the x, y and z axis, respectively. There is no complete matching of the rankings but a general correlation trend exists. Spearman correlations (approximate values at the second decimal point): i) Luo screen: Standard-RIGER = 0.74, Standard-RSA = 0.7, RIGER-RSA = 0.8, ii) Wang screen: Standard-RIGER = 0.65, Standard-RSA = 0.59, RIGER-RSA = 0.76

1. The intersection among the retrieved hits for each screen was calculated. The aim was to have relatively restricted results for greater accuracy, but still a sufficient number of genes for proceeding to next step filtering. There are 1069 common genes among the three screens, which is a proper enough amount to continue with. That is why the intersection was chosen instead of the union for the retrieval of within method hits.

2. Both the intersection and the union of step 1 findings (intersection of hits from all the screens for each method) was investigated. The intersection returned a narrow set of candidates, which are probably the more robust ones.
   RIGER & RSA: *BRCA2, PSMD12, SPRY1*
   RIGER & Standard Method: *SPRY1, FOS, COPB2*
   Standard Method & RSA: *SPRY1*

   It is worth noticing that SPRY1 is the only common gene among all screens and all methods. This will be further considered, when the final list of candidate genes is reported.

   However, the union of the methods was kept for further analysis as it allowed for a broader range of candidates at first step.

Table 4.2: Retrieved genes by each method. Intersection among the 3 screens.

Standard Method	**RSA**	**RIGER**
AKAP8L	BRCA2	BRCA2
ASB2	KCNN3	COPB2
BACH2	NCOR1	CTNNA1
COPB2	POLR2H	EZH1
DNTTIP1	PSMD12	FOS
ERN1	SH3BP4	GSG2
FOS	SPRY1	KCNRG
GUCY1A3	TOP1	MARCKSL1
HDAC9	USP48	MECOM
ISL2		NFKB1
MYBL1		PAX6
NKIRAS1		PSMD12
PLAG1		PSMD3
RAB7L1		RPL30
RAP1A		SMAD1
SPRY1		SON
SUV39H2		SPRY1
TTK		
UBE2I		
WASF3		

Table 4.2 contains the intersection of genes that are retrieved by each method for all three screens.

4.6 Genes with many of their shRNAs ranking below threshold

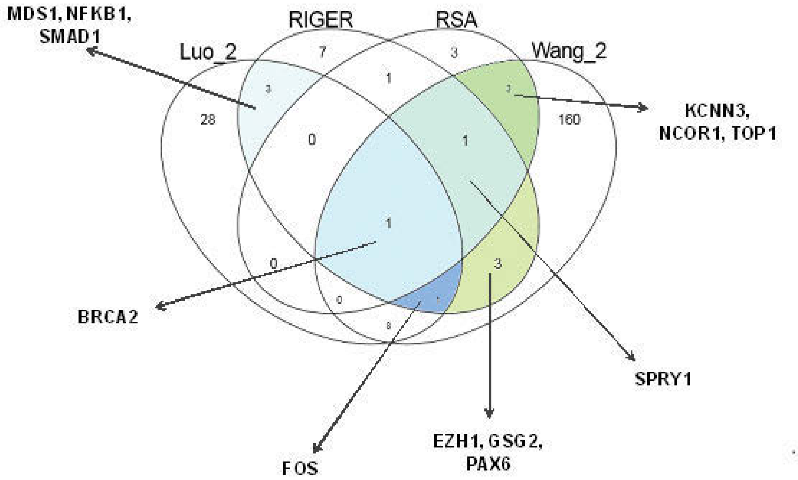


FIGURE 4.9: Intersections of genes of which the shRNAs rank twice below our threshold for the standard method, for both Luo and Wang screens which have multiple shRNAs per gene ($Luo\_2, Wang\_2$), with RIGER and RSA results.

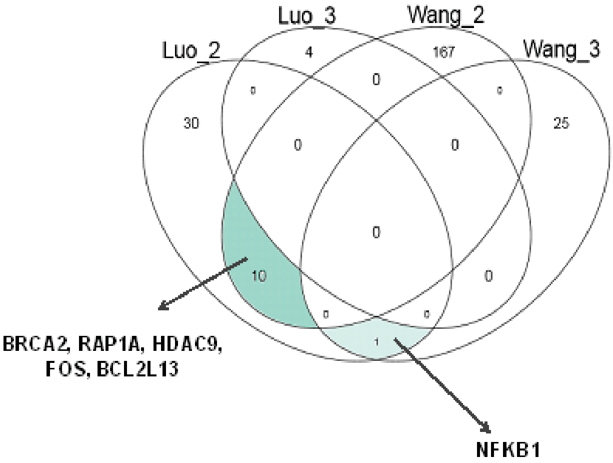


FIGURE 4.10: Intersections between the genes of which the shRNAs rank three times ($Luo\_3, Wang\_3$) or twice ($Luo\_2, Wang\_2$) below our threshold for the standard method, for both Luo and Wang screens which have multiple shRNAs per gene.

Regarding the Luo and Wang screens, some genes were targeted by shRNAs that scored as hits more than once, when the Standard method was applied. On the other hand, the RSA and RIGER methods take as input all the shRNAs but they return only one

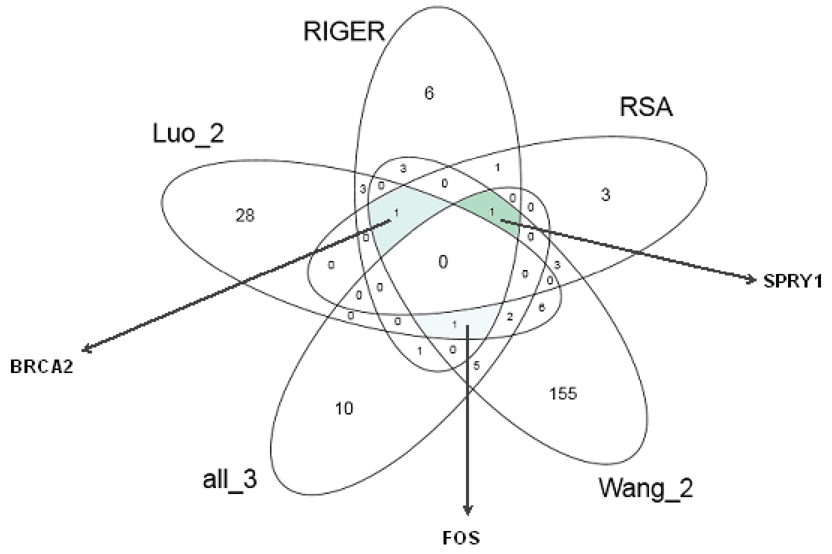


Figure 4.11: Intersections of genes of which two shRNAs against the same genes rank below threshold for the standard method, for both Luo and Wang screens $(Luo\_2, Wang\_2)$, with RIGER and RSA results. Additionally to Figure 4.9, the intersections with the hit genes for the standard method and for the three screens Luo, Steckel and Wang is depicted, considering only the best ranking shRNA.

p-value per gene. So, each significant gene was retrieved only once after applying the respective threshold criteria. Figures 4.9, 4.10 and 4.11 present with Venn diagrams the intersections between the methods for the Luo and Wang screens, when multiple shRNAs that target the same gene ranked below threshold by the Standard method. From the above, it seems that some genes are more potent candidates. These are: BCL2L13, BRCA2, EZH1, FOS, GSG2, HDAC9, KCNN3, MECOM, NCOR1, NFKB1, PAX6, RAP1A, SMAD1 and TOP1. That is because they intersect among the methods. The significance of each intersection was calculated using the *phyper* R function. This function decides on the significance of the results retrieved by two different methods, by using a hypergeometric distribution test on the intersection of methods. Table 4.3 shows the results of this test. Some intersections were significant while others weren't. The additional intersection Wang2-Luo2, not shown in the table, was also calculated. The respective p-value was ~ 0.124 which is far from the significance level.

Selection of the genes from the significant intersections concluded to 22 genes (out of the initial 40, contained in the union of Standard, RIGER and RSA methods - Table 4.2): FOS(4), SPRY1(3), BRCA2(3), RAP1A(2), NFKB1(2), HDAC9(2), TTK(1), TOP1(1), SUV39H2(1), SON(1), SMAD1(1), PAX6(1), NCOR1(1), MYBL1(1), MECOM(1), KCNN3(1), GSG2(1), EZH1(1), DNTTIP1(1), CTNNA1(1), COPB2(1), BACH2(1). The frequency of significant intersection is indicated in the parenthesis. Additional methods were needed to support the validity of this initially retrieved set of genes.

Table 4.3: Significance of intersections among the 3 screens for the Standard, the RIGER and the RSA methods

	Luo-Steckel-Wang (best shRNA)	**RSA**	**RIGER**
Luo2	0.03792517	0.2977	0.00029
Luo3	-	-	0.062
Wang2	0.002453423	6.51 e-8	0.0476
Wang3	0.3916075	-	0.007

4.7 Results from RNAiCut

The results after application of the RNAiCut method are shown in Table 4.4 in juxtaposition with the results from the Standard method.

Table 4.4: Comparison of retrieved hit genes between RNAiCut and Standard ranking computations, applied on the 1069 common genes only

Screen	**N_o of genes: RNAiCut**	**N_o of genes: Computations**	**Intersection**
Luo	101	223	22
Steckel	217	Stringent (Z=-2): 35	31
		Relaxed (Z=-1): 126	98
Wang	136	685	136

- For the Luo screen, the threshold was initially set too low by RNAiCut: only four genes out of 1069 were selected as true positives. I contacted the author of the respective paper, Prof Irene Kaplow, and she indeed admitted that sometimes the algorithm gets trapped in local minima. She suggested to look for the next local minima which in this case is met at the 101^{st} hit. From the original analysis, a set of 223 genes was retrieved, including 22 of the aforementioned 101. This difference is raising questions but at least assigns highest confidence to these 22 genes.

- For the Steckel screen, 217 genes were selected by RNAiCut, while 35 genes (stringent criteria) and 126 genes (relaxed criteria) were selected by the Standard method. Table 4.4 shows that most of the genes coming from computations were included in the set of genes retrieved by RNAiCut. It is interesting that RNAiCut stresses even more genes than the ones retrieved by applying the relaxed criteria of z-score $<$ -1.

- For the Wang screen, 136 genes were detected by RNAiCut, whereas 685 genes were detected by the Standard method. The larger set contained all the genes of the smaller set.

Table 4.5: Comparison of retrieved hit genes between RNAiCut and the three basic ranking methods

intersection	Standard method	RIGER	RSA
Luo	ERN1, BACH2, PLAG1, COPB2, MYBL1, WASF3 HDAC9, FOS, SUV39H2 NKIRAS1	BRCA2, CTNNA1, SMAD1, COPB2, FOS MDS1, NFKB1	BRCA2
Steckel	All	FOS, SON, SPRY1 BRCA2, COPB2	BRCA2, SPRY1
Wang	BACH2, SUV39H2, TTK HDAC9, FOS, RAB7L1 ISL2, DNTTIP1, SPRY1 COPB2	GSG2, NFKB1, PAX6 SON, FOS, COPB2 EZH1, SPRY1	SPRY1, TOP1

The genes retrieved by RNAiCut were investigated and compared with the genes that were retrieved by each of the three basic ranking methods. Almost all the genes that were detected by the Standard method, RIGER and RSA were below the threshold that RNAiCut imposed. Moreover, BCL2L13, a gene that is present in the intersections of two screens (see Venn diagrams), but not of three screens, was also selected by RNAiCut. On the one hand this high intersection was a positive result, since it implied that RankSLP was heading towards the right direction. On the other hand, RNAiCut did not provide any significant additional information at the final selection of hit candidates.

4.8 Incorporation of the genes retrieved by rank aggregation

The rank aggregation method was applied by using the *R* function *RankAggreg()* (see details in section 3.4). It was applied on the ranks of the Standard method for the three screens. For detecting the hit genes, significance thresholds were imposed. However, for the case of RankAggreg, the whole ranked list is needed as input and the algorithm requires the use of only one shRNA per gene. Therefore, in the case of Luo and Wang screens, the best ranking shRNA was used, for consistency with Standard method. In this part, the findings of *RankAggreg R* function are exhibited. As explained in the *Methods* section, the top 10% of the genes that would rank high in a combined *super*-list were collected.

Two heuristics towards the generation of the list with the minimum Spearman distance to all three lists, were compared: the Genetic Algorithm (GA) and the Cross Entropy Monte Carlo (CE). The Genetic Algorithm did not converge, but two genes, REL and AKT1 were selected very often, with different parameter choice.

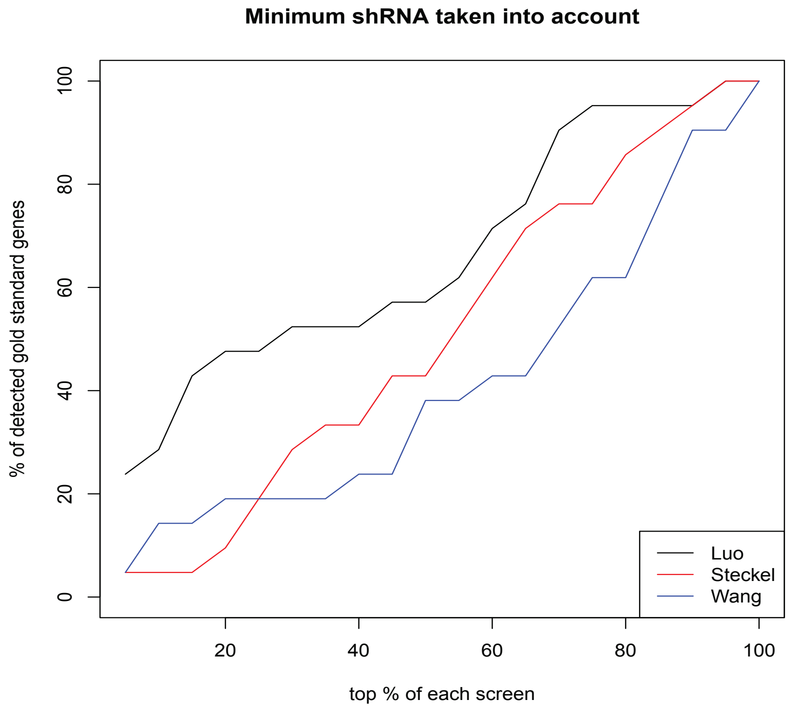


FIGURE 4.12: Frequency of retrieved GSGs at the top % of each ranked list, as calculated by the Standard method. The best ranking shRNA was considered in Luo and Wang screens. It is clear that Luo screen ranking outperforms the other two.

The CE algorithm converged and its findings were further used. CE is also preferred by the developers of the method, as it performs better than GA on real data [59]. The chosen distance measure was the Spearman footrule distance. The rarity parameter ρ used in updating the cell probabilities was set to 0.01.

RankSLP assigned importance coefficients on the lists based on how well they rank the GSGs at their top 10% positions. Figure 4.12 clearly shows that, in these terms, the Luo screen is of better quality. For proper calculation of the importance coefficients that should be assigned to each screen, the fractions of GSGs that are retrieved at different cutting thresholds starting from the top of each ranked screens were calculated. These fractions are measured by dividing the number of GSGs that are ranked at top 10% positions, by the number of GSGs that are part of the common genes among all three screens (21). The percentages are $\frac{6}{21}$, $\frac{1}{21}$ and $\frac{3}{21}$ for the Luo, Steckel and Wang screens respectively.

RankSLP looked for a proper multiplier *mult* to standardize these percentages so that they sum up to one. Thus, the following simple equation model was fitted (4.1):

$$mult = \frac{1}{perc.luo + perc.steckel + perc.wang} \tag{4.1}$$

$$weight\_Luo = perc.luo * mult = 0.6$$

$$weight\_Steckel = perc.steckel * mult = 0.1$$

$$weight\_Wang = perc.wang * mult = 0.3,$$

where $perc.luo, perc.steckel$ and $perc.wang$ are the importance weights, corresponding to the percentages of GSGs that were retrieved at the top 10% positions of each ranking.

The resulting genes that are common with the identified hits of sections 4.5 and 4.6 are COPB2, NCOR1 and SPRY1 (see Table 4.6). RankSLP also experimented with the significant genes when aggregating the top 5% of the ranked input lists. The above genes were retrieved as well, indicating that these three are quite robust both to the ranking method and to the screen.

Table 4.6: Intersections of weighted top 5% and top 10% RankAggreg results, with Standard, RIGER and RSA methods

Method	**Common genes, top 5% or 10%**
Standard method (relaxed criteria)	COPB2, SPRY1
RIGER	COPB2, SPRY1
RSA	NCOR1, SPRY1

4.9 First hit genes selection and pathways enrichment

At this stage some first results were already retrieved. Before proceeding to evaluation, the pathways of the hit genes were investigated, in order to estimate if the previous analysis was correctly directed. To this aim, the genes that were confirmed by at least two of the already examined methods were selected. The 25 resulting genes were searched in five publicly available databases: i) WikiPathways [84], ii) GeneCards [85], iii) KEGG, release 71.0, July 1 2014, iv)PubMed Gene, accessed in May 2014 and v) REACTOME pathway database[5]. The main pathways in which the hit genes participate are shown in table 4.7. The majority are cancer pathways. This was a proof that the followed procedure leads to meaningful findings.

4.10 Comparison with literature genes

In this section, a comparison of the current findings with the KRAS synthetic lethal partners that are discussed in existing literature is provided (GSGs - see subsection

---

[5] The accessed databases following the text order are `http://www.wikipathways.org`, `www.genecards.org`, `http://www.genome.jp/kegg`, `http://www.ncbi.nlm.nih.gov/gene`, `http://www.reactome.org/`

Table 4.7: Pathways in which the hit genes participate. The genes are ordered alphabetically.

BACH2	lymphocyte signaling
BCL2L13	legionellosis
BRCA2	cell cycle, HR (Homologous recombination), DNA repair
COPB2	membrane trafficking
CTNNA1	hippo pathway, adherence junction
DNTTIP1	no information found
ERN1	Alzheimer disease, unfolded protein response
EZH1	gene expression
FOS	MAPK/ERK, Wnt signaling, TNF
GSG2	no information found
HDAC9	histone deacetylation, viral carcinogenesis
MECOM	MAPK/ERK
MYBL1	HTLV-I infection, IL4-mediated signaling event
NCOR1	notch-delta pathway, signaling by Erbb4
NFKB1	MAPK/ERK, NF-kappa B, PI3k/Akt/mTOR, Viral carcinogenesis
PAX6	CDC42 signaling events, maturity onset diabetes of the young, regulation of gene expression in beta cells
PSMD12	proteasome
RAP1A	chemokine signaling pathway, focal adhesion, IL-3 Signaling Pathway, MAPK/ERK, long-term potentiation
SMAD1	hippo pathway, angiogenesis, TGF-beta signaling, transcriptional misregulation in cancer
SON	NCAM signaling for neurite out growth, oxytocin signaling pathway
SPRY1	JAK/STAT signaling pathway, signaling by EGFR
SUV39H2	lysine degradation
TOP1	caspace cascade in apoptosis
TTK	RB/E2f pathway in cancer, cell signaling, cell cycle, checkpoint control
UBE2I	cell cycle, meiotic synapsis, SUMOylation, TNF-alpha/NFKB, ubiquitin mediated proteolysis

4.1.2). To this aim, the comparison was done with the GSGs that are retrieved from literature only. It is crucial that this comparison is done with the external to the three used screens' data only, in order to avoid overestimation of the ranking methods used. The literature genes that are supported by external sources are 28 in total. Out of these, 21 intersect with the common genes among the three screens. RankSLP detected only two of them, NFKB1 and SMAD1. This limited overlap, made me investigate the ranks of the GSGs that were retrieved by the Standard, RIGER and RSA methods, described in the Data and Methods Chapter. In the cases where more than one shRNAs target the same gene, its rank was the best rank of its shRNAs. Simple consensus metrics, including the minimum, maximum, average and median of the individual screen ranks were calculated across the three screens for each gene. Figures 4.13 and 4.14, illustrate these consensus ranks and the percentage of GSGs that are retrieved in the first quantile (top 25%) of ranks.

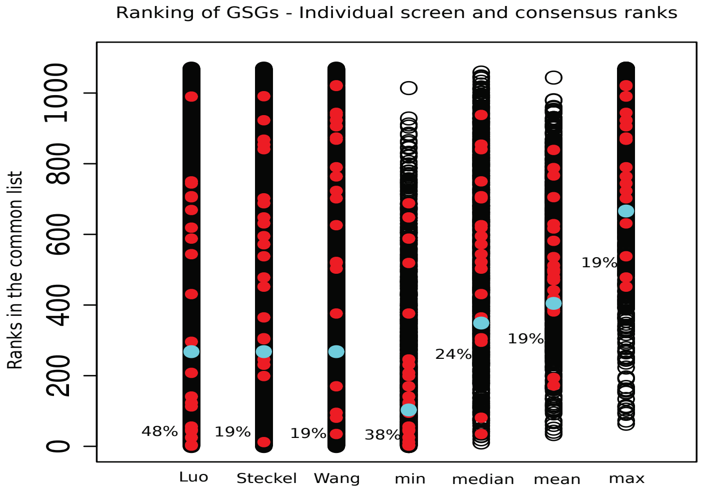


FIGURE 4.13: Consensus ranks of the common genes across the Luo, Steckel and Wang screens. The ranks are plotted in black dots. The red dots are the GSGs and the cyan dot indicates the first quantile. The x axis shows the consensus metric that is utilized in each bar. At the bottom of the bar is shown the percentage of GSGs that lay in the first quantile.

The coverage of GSGs at the first quantile is varying, exhibiting highest scores for the Standard method on the Luo and Wang screens and for the RSA method. Judgement of which of the methods (Standard method, RIGER and RSA) is best based on the ranks that it assigned to GSGs was not a trivial task. There is not a clear pattern and the number of GSGs is small (21). This is further discussed in the next Chapter. Overall though, the GSGs rank relatively well.

It has to be noticed that three of the detected genes, COPB2, PSMD3 and PSMD12 (the two latter are proteasome components) were confirmed internally by Luo and Wang screens. This is not as strong evidence as for the NFKB1 and SMAD1 genes, which were confirmed by external data.

4.11 Comparison with high throughput drug screens

In an RNAi screen, a gene of interest is knocked down (silenced) by an siRNA or shRNA that has complementary sequence to it. In the same way, in a *compound inhibition screen* or *drug screen*, a drug targets a gene of interest in order to silence it. For KRAS mutant cell lines and in both RNAi and drug screens, if a KRAS SLP is targeted and knocked

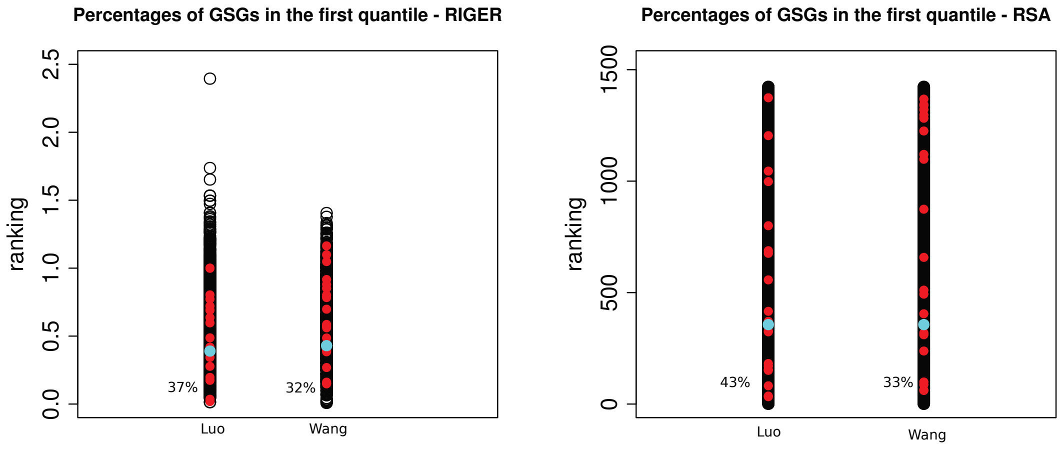


FIGURE 4.14: The figures depict the RIGER (left) and RSA (right) ranks of the genes that are common between the Luo and Wang screens remember, RIGER and RSA can only be applied in these two screens because they require multiple shRNAs per gene). The ranks are plotted in black dots. The red dots are the GSGs and the cyan dot indicates the first quantile. The x axis shows the screen of which the rankings are represented by each bar. At the bottom of the bar there is the percentage of GSGs that lay in the first quantile.

down (KD) by an RNAi inhibitor or drug D respectively, the cell dies. So, following the reverse path, our hypothesis is that: If for two KRAS mutant cell lines being targeted by drug D and shRNA corresponding to gene KD the same lethal phenotype is observed, then drug D targets gene KD. This hypothesis is shown in Figure 4.15, right hand side.

TABLE 4.8: Chemical Genomics Screens

The field *Year* corresponds to when the respective study took place. CCLE, GDSC and NCI60 datasets are often being updated. The versions of June 2013 were used.

Screen Name	**Year**	N_o **of Compounds**	**Cell Lines**
NCI60	1990 - Present	~20000	59
GDSC	2012	138	714
CCLE	2012	24	504
Steckel Drug	2012	108	2

Based on this hypothesis, four external drug screen datasets were analyzed in order to compare and confirm the findings from the RNAi screens. These are summarized in Table 4.8

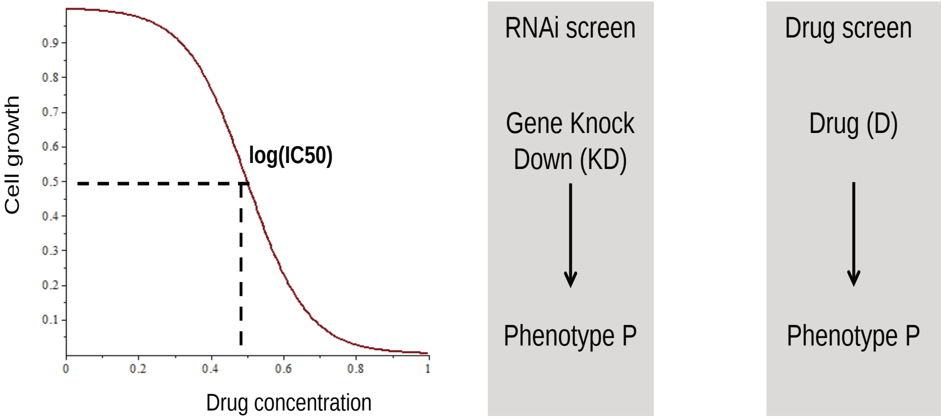


FIGURE 4.15: Left: Drug response curve. The viability of the tumor cell line is dropping as the drug concentration is increasing. A 50% decrease in tumour growth corresponds to the IC50 drug dosage (dashed line). Right: Our hypothesis: If P is similar to P, then we suppose that drug D targets gene KD

4.11.1 Dataset retrieval

In this part, a general description of each drug dataset is provided.
1) Cancer Cell Line Encyclopedia (CCLE) [49]: This project was initiated in 2012 by Broad Institute and consists of an effort to systematically characterize genetically a large panel of human cancer cell lines, in terms of their mutation, DNA CNV, SNP e.t.c. A part of this project is the drug response testing of 24 anticancer drugs on these cell lines. In total, 1000 cell lines are examined and the mutation status of $\sim$ 1600 genes is captured. The drug response data were downloaded in May 2013 from the *CCLE* database[6].

2) Genomics of Drug Sensitivity in Cancer (GDSC) dataset [86]: This is an effort of the Sanger Institute towards the detection of cancer therapeutic biomarkers, through the examination of the responses of cancer cells to specific drugs. To this aim, 700 cancer cell lines are tested for their sensitivity/resistance against 138 chemical compounds. Overall, around $\sim$ 75000 experiments have been conducted. For the current analysis, the resource named "cell line drug sensitivity, mutations and tissue type" was downloaded by the *GDSC* database[7], in April 2013.

3) National Cancer Institute dataset (NCI60): This data is part of the Developmental Therapeutics program which was initiated in late 1908's by the National Institute of

---

[6] `http://www.broadinstitute.org/ccle`
[7] `http://www.cancerrxgene.org/downloads/`

Health (NIH) in Bethesda - US, and it is still ongoing [87][8]. The number 60 comes because of the effects on 60 tumour cell lines. However, complete data allowing further analysis exist only for the 59 of them. The datasets are accessible via the CellMiner portal [88], [89]. The cell line-wise normalized drug response data (DTP) were downloaded from this site. The respective dataset contains 20,502 compounds and their effects on 59 human cancer cell lines.

4) Steckel Drug: The fourth *Steckel Drug* screen, is performed by the same group of scientists that conducted the *Steckel* RNAi screen [26] and is freely accessible online from the publication website.

4.11.2 CCLE, GDSC and NCI60 screens

Three of the four screens, namely *CCLE*, *GDSC* and *NCI60* contain information about each cell line's sensitivity to each drug. This is captured by the **IC50** value: This value represents the compound dosage that reduces the natural tumour growth of the cell to 50% in 48 hours. The drug response curve shows the decrease in tumour growth with respect to a drug's concentration. The shape is usually sigmoid and an example is shown in Figure 4.15 along with the IC50 value.

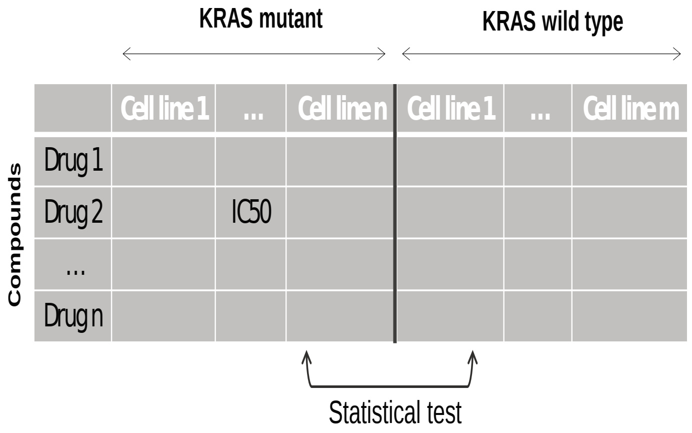


FIGURE 4.16: The KRAS mutant cell lines were separated from the KRAS wild type. The wilcoxon rank sum test was applied with alternative *less*. The H_0 assumes that there is no differential effect of the drug between the two cell line categories.

In this case a simple statistical analysis between KRAS mutant and KRAS wild type cell lines was performed. The mutation status for CCLE data were provided by the CCLE

---

[8] `http://dtp.nci.nih.gov/index.html`

portal. Only the mutations that led to a protein change (no UTR's e.t.c.) were considered. The GODS and NCI60 mutations were retrieved from additional files provided together with the data. Extra searching was required in few cases; these additional mutation data were obtained from the Sanger Institute Catalogue Of Somatic Mutations In Cancer[9] (COSMIC) [90]. The performed statistical analysis involves the following four steps:
1) Initially all the IC50 values were converted to micromolars (uM), which is the most common unit for drug concentrations.
2) Afterwards, the data were subjected to double normalization, drug-wise and cell line-wise, in order to avoid all possible biases in further analysis.
3) Finally, the *Wilcoxon rank sum* test was applied to the datasets, as shown in Figure 4.16. The H_o assumes that there is no differential drug effect between the KRAS mutant and the KRAS wild type cell lines.
The alternative was set to *less*.
4) The significant drugs were the ones that had a p-value < 0.05. By rejecting the H_o at the 95% significance level, it means that there exists just a 5% probability that the observed less viability of the mutant cells versus the wild type is attributed to chance.

4.11.3 Steckel Drug screen

In the case of the Steckel drug dataset, IC50 values are not available. The data consist of four different concentrations of each drug and the viability ratio wt/mut.

In order to detect the significant drugs, based on their effect on KRAS mutants viability, the following steps were performed:
1) Deletion of negative control rows.
2) Inversion of the viability ratio so as to have mut/wt viability R.
3) Normalization over the negative controls $R' = (R - \mu(nc))/\sigma(nc)$
4) Selection of the drugs with $R' < 0$, because the interesting drugs are the ones that lead to a smaller viability for mutants than for wild types.

4.11.4 Drug-Target matching

In previous sections, the selection scheme of significant drugs is explained. But how can these results be compared to the results coming from the RNAi screen analysis? To this aim, STITCH v3 [91], a fully up-to-date database that contains all known drug target

---

[9] `http://www.sanger.ac.uk/cosmicwere`

relationships for all species integrated from other databases, was used. The database is freely accessible online at `http://stitch.embl.de/`. The underlying structure of STITCH is a graph with nodes representing compounds or genes and edges representing their connections (see Figure 4.17). The color and width of each edge correspond to the source from which the specific connection is provided and to the confidence that this connection really exists, respectively.

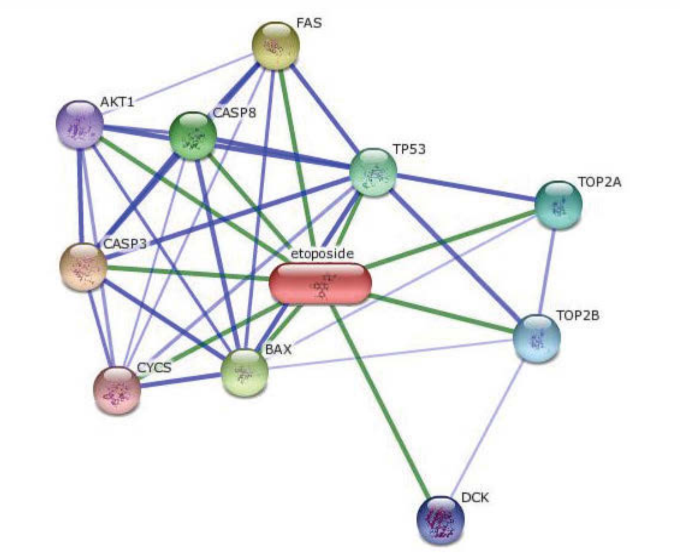


FIGURE 4.17: An example screenshot of a search for the compound *'etoposide'* in the STITCH database.

The frequency of the number of compound-target pairs in STITCH with respect to the confidence threshold is presented in Figure 4.18. For the current analysis, STITCH drug-target pairs with a relatively "high" confidence threshold of at least 70% were considered, as suggested by the developers of the database. Since STITCH provides protein identifiers, matching with the respective gene identifiers was performed through BioMart interrogation[10], using the R Bioconductor package *bioMart*.

For each of the drug datasets, RankSLP matched their provided identifier with the ones provided in Stitch. For all the drug datasets the matching was done using aliases. For the CCLE, GDSC and Steckel Drug datasets, this was the only provided identifier. For the NCI60 dataset, the SMILES codes of some of the tested compounds were provided. The matching with STITCH was first attempted through InchiKeys. The SMILES-to-InchiKey conversion was done using the Open Babel Package[11] (accessed in June 2013). However, the SMILES codes were provided as a seperateset list (IC50) to the drug response data. Only 2,530 of these SMILES codes were accompanied by the respective

---

[10] `http://www.biomart.org/`

[11] `http://openbabel.sourceforge.net/`

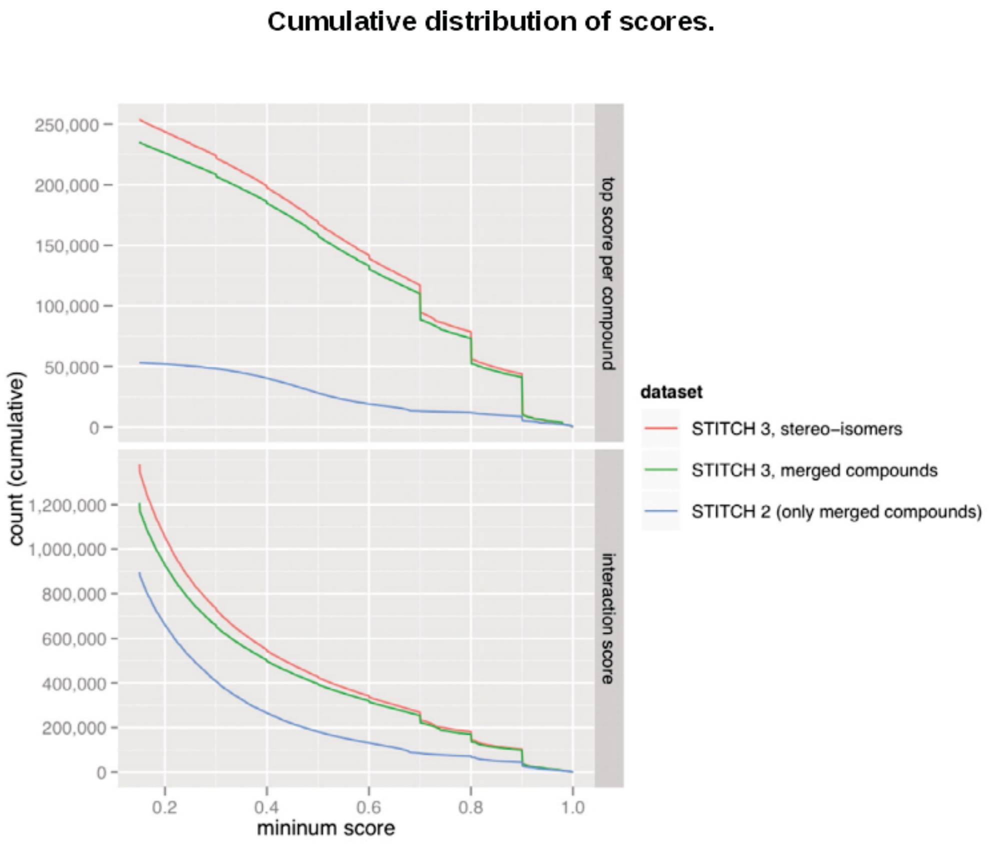


FIGURE 4.18: Cumulative distribution of scores. For each confidence score cutoff, the number of chemicals (top) and protein–chemical interactions (bottom) that have at least this confidence score in the human protein–chemical network. For example, there are 110,000 chemicals with a high-confidence interaction (score at least 0.7). Note that interactions with confidence scores below 0.15 are not stored in STITCH. Steps in the data correspond to large numbers of compounds that have a maximum score in manually curated databases or the ChEMBL database (with different confidence levels). Both the figure and the caption are taken from the original paper [91].

alias, which is required for matching with the IC50s. On the other hand, the NCI60 compound aliases that can match with STITCH (using the alias only) are 2,682. That is why the matching based on aliases was further used. Unfortunately, since the aliases are not unique identifiers, this resulted in an incomplete matching of the compound names with STITCH. The matching algorithm has $O(nk)$ complexity, where n is the amount of drugs that pass the 70% threshold criteria in STITCH and k the amount of compounds in each drug dataset. Considering the vast amount of drugs with known targets in STITCH, this procedure takes much time(approximately one hour for CCLE which is the smallest drug dataset and three to four hours for NCI60 which is the largest drug dataset).

4.11.5 Identified genes from the analysis of drug response data

The drug screen analysis was completed by concluding in the emerging gene hits. A crucial assumption that was made at this stage, was that if a drug-target pair is detected, it is considered as valid, without taking into account the effects of *polypharmacology*. This is a scenario that was followed in order to decrease the problem's complexity. To be totally realistic, we should have considered the effects of the selected drug on other target genes and potential interactions among them should have been considered. This assumption is further discussed in the Discussion Chapter.

Table 4.9 provides the results in a summarized way.

Table 4.9: Drug screen analysis results

Screen	Hits also confirmed by RNAi screens
CCLE	BRCA2, HDAC9, NFKB1, TOP1
GDSC	BRCA2, ERN1, FOS, NFKB1, TOP1
NCI60	FOS, NFKB1
Steckel Drug	BRCA2, FOS, HDAC9, NCOR1, NFKB1, PAX6, TOP1

Some additional notes on Table 4.9 are:

- Regarding the CCLE dataset, the targets of the 24 screened drugs are provided under the official CCLE portal[12]. Only one protein target per drug is given and we suppose that this is the *intended* target. Table 4.10 shows the differential drugs between KRAS mutant and KRAS wild type cells and their intended targets in the CCLE database. As it can be noticed, there is no exact match between these genes and the ones that were retrieved by taking all the targets of a drug using a

Table 4.10: CCLE: Significant compounds and their intended targets, as provided by the official CCLE portal `http://www.broadinstitute.org/ccle`

Compound	Intended target protein
17-AAG	HSP90
AZD0530	MEK
Irinotecan	TOP2
Nilotinib	ABL
Paclitaxel	TUBB1
PD-0325901	MEK
ZD-6474	EGFR

---

[12] `http://www.broadinstitute.org/ccle`

70% threshold, as described in section 4.11.4. However, there is a general trend for silencing of the MAPK/ERK pathway (MEK, EGFR). This is in accordance with the detection of the FOS oncogene as hit. In addition, TOP1 is detected as target using STITCH analysis, instead of TOP2.

- Regarding the NCI60 dataset, only few compounds had a significant differential effect between the cancer cell lines. Attention was paid to that; a larger number, given the amount of compounds (around 20,000) that were tested. Searched for supporting evidence I came across the 2012 paper by Burkard M. E. [92], where the authors show that there are few differential drugs in the NCI60 database.

The SPRY1 gene, the only common hit between the three main used methods for RNAi screen ranking (see section 4.5), was not detected by the drug screens. This case was investigated and it was found that indeed, no drugs that target SPRY1 are screened in any of the four examined compound inhibition datasets.

4.12 Network enrichment

In this section, the position of the hit genes in the human protein network and their connections are investigated. The hypothesis is that strong network connections of the hit genes and possible interactions with KRAS would help to verify the findings of section 4.9. To this aim, the retrieved genes and their containing modules were examined using WebGestalt. WebGestalt stands for“WEB-based GEne SeT AnaLysis Toolkit” and it is a freely accessible online tool for gene enrichment analysis, provided by the Vanderbilt University [93], [94]. It is mainly used for the analysis of the significant genes (i.e. differentially expressed genes) resulting from high throughput genetic datasets, like functional genomics and proteomics. WebGestalt contains information from different public databases, and provides a unique interface to all of them, facilitating and speeding up the expanded analysis of genetic data.

In RankSLP, homo_sapiens was selected as the organism of interest. The gene ID type was hsapiens__gene_symbol. The enrichment analysis methods that were used are *Pathway Commons*, *Gene Ontology* terms (GO) and *Protein Interaction Network Module.* The reference set was hsapiens__genome and the statistical method was hypergeometric. Further multiple test correction was performed using the Benjamin-Hochberg procedure (BH). The significance level was set to top 10 and the minimum number of genes for a category was three. Using the above parameters, significant modules were highlighted. The selected modules were mapped into Cytoscape, version v.3.2[13], using the STRING

---

[13] `http://www.cytoscape.org`

network connections as of the March 2014 version. No confidence threshold was imposed on the edges of the network. The visualization was a five-step procedure:

1) For each of the modules, its consisting genes were detected and highlighted.
2) Their first neighbors were retrieved (the undirected version of the network was used for that)
3) A subnetwork, containing all the nodes and edges that are part of steps 1 or 2, was created.
4) The Gold Standard Genes along with the KRAS gene were also highlighted in the final network.
5) The network was re-arranged to ameliorate its readability. The genes that were end nodes without further connections were deleted for simplicity, except if they belong to the ones retrieved in steps 1-4 or have some known important cancer-related function.

4.12.0.1 Pathway Commons analysis

One of the analysis available by WebGestalt is Pathway Commons [95]. It highlights the common cell signaling and metabolic pathways in which the retrieved genes participate, providing also information about gene-gene interactions. One set of RankSLP's hit genes participates in the TNF alpha and the NF-kB pathways (Figure 4.19), which are known to be cancer-related, and another set is enriched in nine different pathways (Figure 4.20). In the first case KRAS was not even present in the network. Many of the hit genes, like HDAC9, NCOR1, NFKB1, PSMD3 and PAX6 are quite central with many connections. This means that many pathways can be affected by their silencing, and thus one should be very careful when knocking any of them down.

In the second case, CTNNA1, FOS, RAP1A, SPRY1, TOP1 and UBE2I exhibit a high connectivity as well, along with the aforementioned NCOR1, NFKB1, PSMD3 and PAX6. Three of our GSGs, are also present: MAP3K7, NFKB1 and SMAD1. The presence of KRAS is noticeable and its direct connectivity with CTNNA1 makes this gene a stronger candidate. Hence, if CTNNA1 is knocked down, the signal will directly be transmitted to KRAS, without intermediate steps that may potentially make it alter or fade.

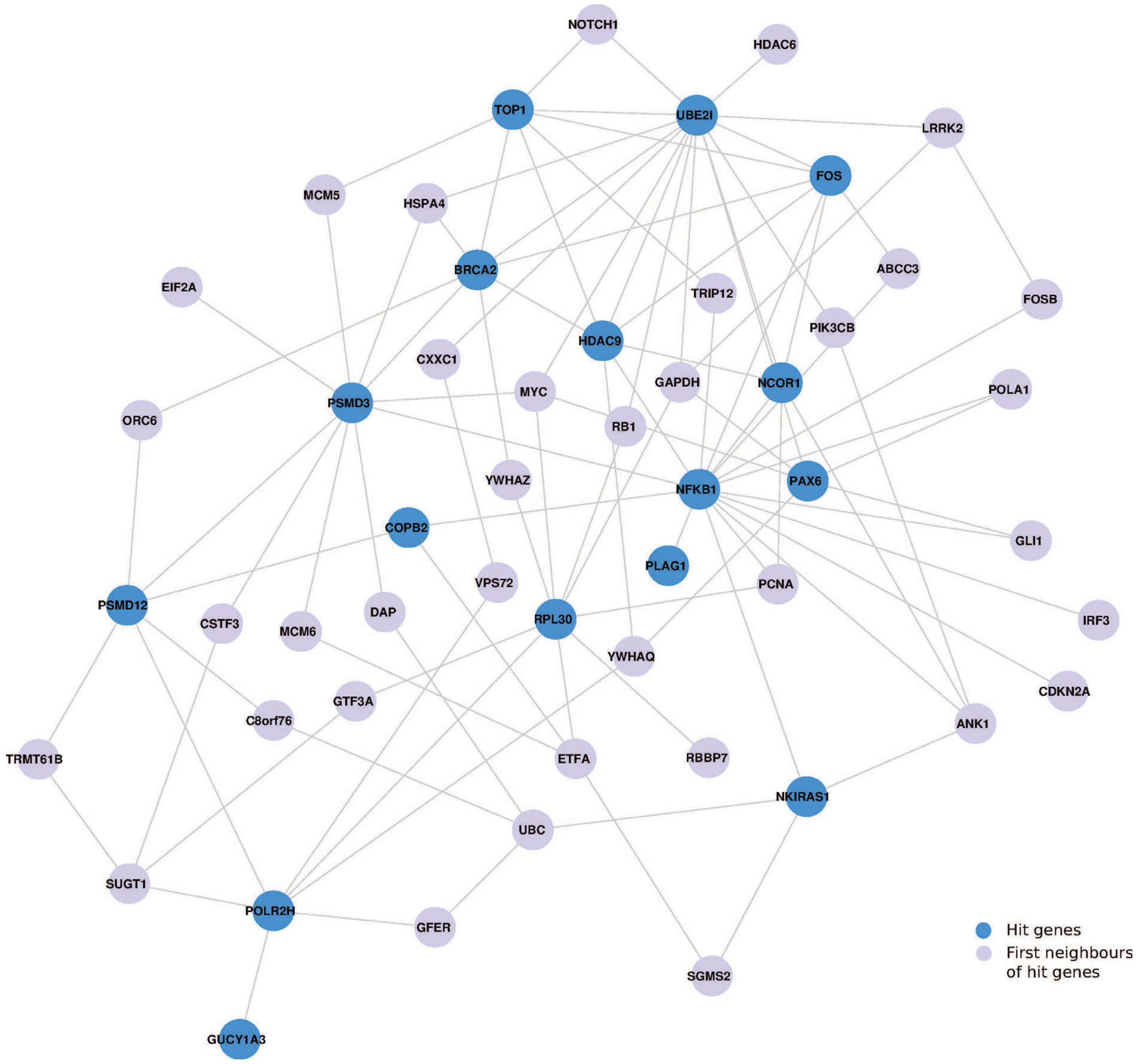


Figure 4.19: TNF alpha/NF-kB pathways: Blue nodes are the hit genes and purple nodes are their first neighbors.

4.12.0.2 Gene Ontology analysis

Another interesting analysis that WebGestalt offers, is protein grouping based on common Gene Ontology (GO) term enrichment. The GO project is a collaborative effort towards consistent descriptions of gene products [96]. It is widely used as a common reference by the scientific community, thus it was interesting to show some of its results on the hits detected by RankSLP. Three example GO categories, with three or more hit genes enriched, are depicted in Figures 4.21 and 4.22. The rest enriched GO categories, having significance (adjusted p-value) of magnitude 10^{-2} to 10^{-4} are presented in Table 4.11.

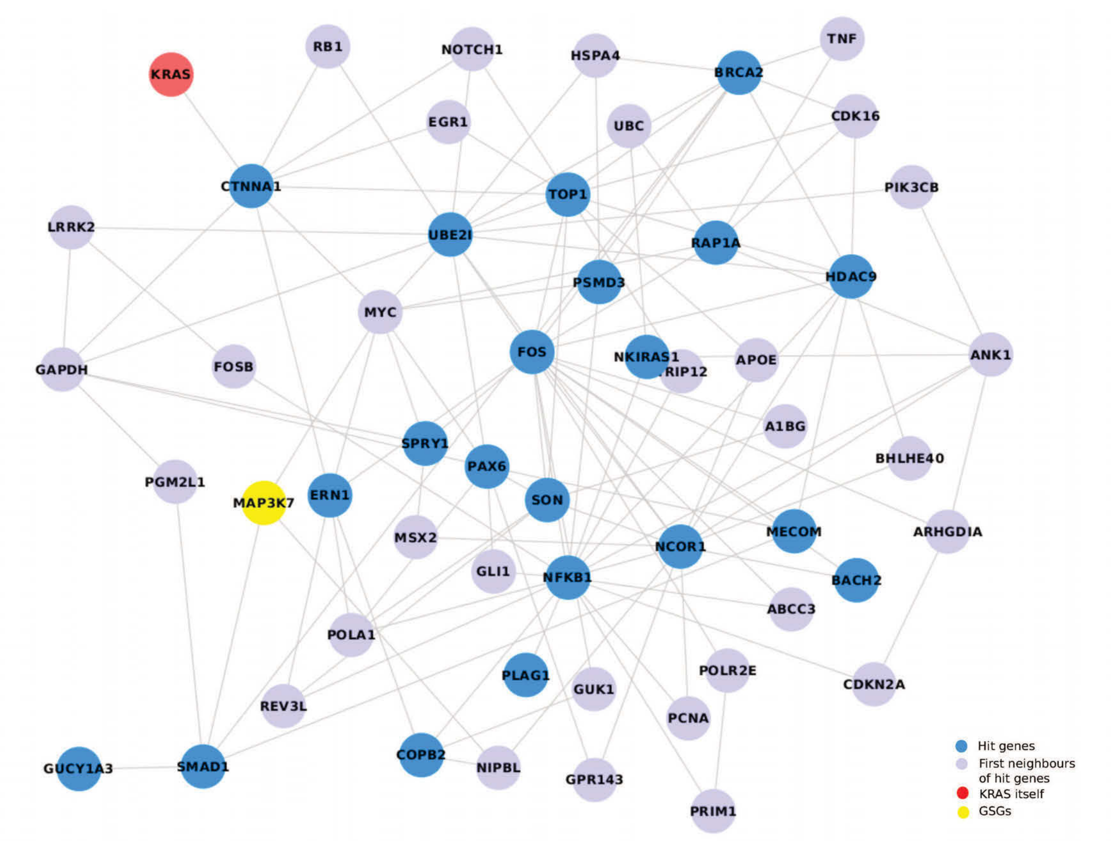


FIGURE 4.20: The highlighted in blue genes participate in nine pathways, having approximately the same significance level ($e-05$): 1. tIFNgamma pathway, 2. Thrombinprotease-activated receptor (PAR) pathway, 3. PDGF receptor signaling network, 4. IL5-mediated signaling events, 5. GMCSF-mediated signaling events, 6. ErbB receptor signaling network, 7. Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling, 8. Internalization of ErbB1 and 9. Plasma membrane estrogen receptor signaling. Blue nodes are the hit genes and purple nodes are their first neighbors. Yellow nodes are the GSGs. NFKB1 and SMAD1 are GSGs as well, but they are also hit genes, thus they are painted in blue.

4.12.0.3 Enrichment having as basis human kinases

The GO categories that were retrieved by the previous analysis, using as reference the whole human genome, are very general. To create a more specific GO enrichment outcome, the human kinases were used as basis. In the Wang screen [27], the targeted genes are "known human cancer genes" and "protein kinases". RankSLP's integrative analysis is applied on the genes that lay in the intersection of the three screens. Thus, they consist of a subcategory of the genes screened by Wang *et.* al. The "known cancer genes" are not a specific set of genes but the human protein kinases are. Hence, the latter were used as a reference aiming at a more informative GO anlysis. A set of 620 human kinases was downloaded from the Human Kinome Project[14] [97], on July

---

[14] http://kinase.com/human/kinome/

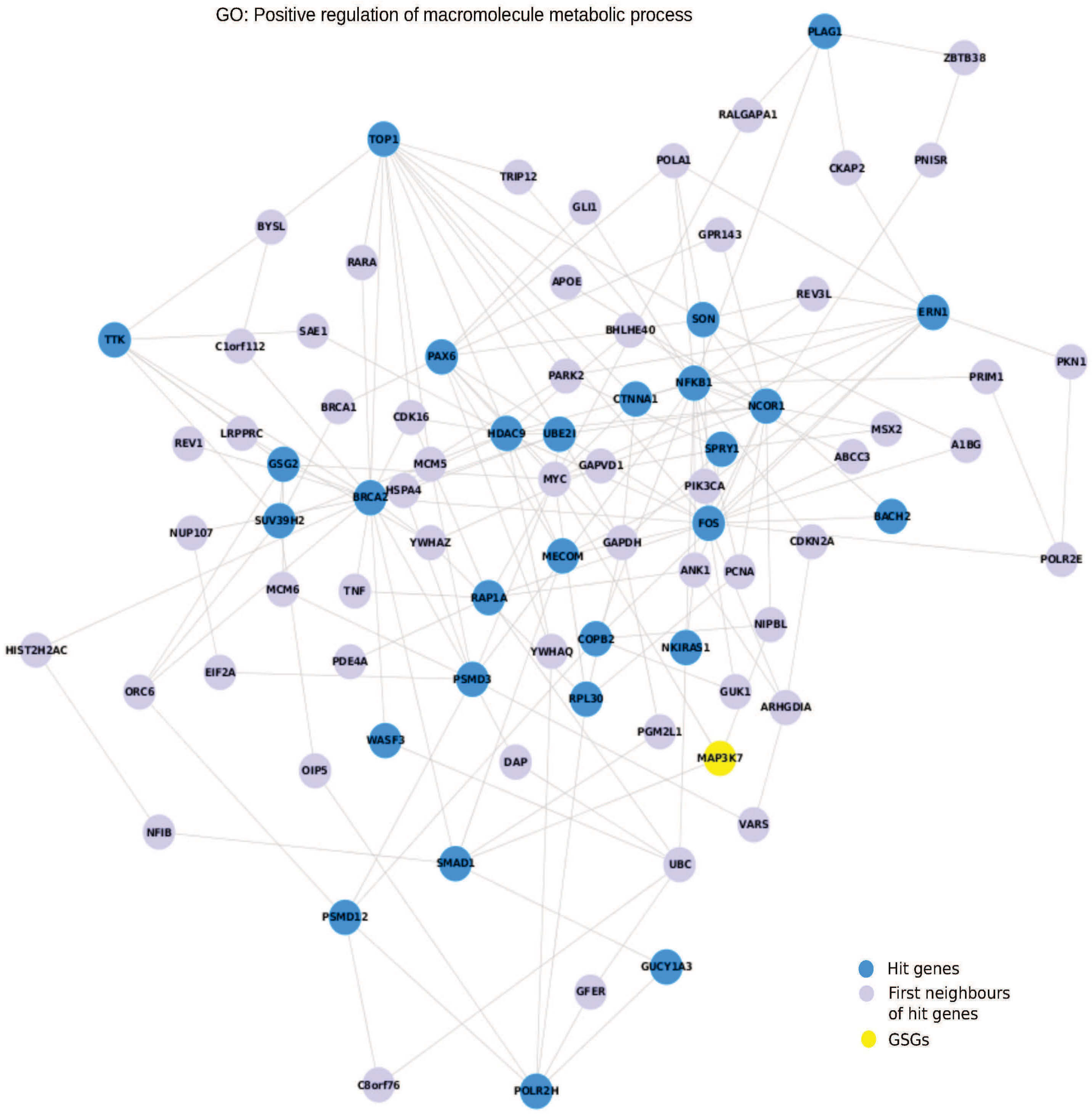


FIGURE 4.21: Blue nodes are the hit genes and purple nodes are their first neighbors. Yellow nodes are the GSGs.

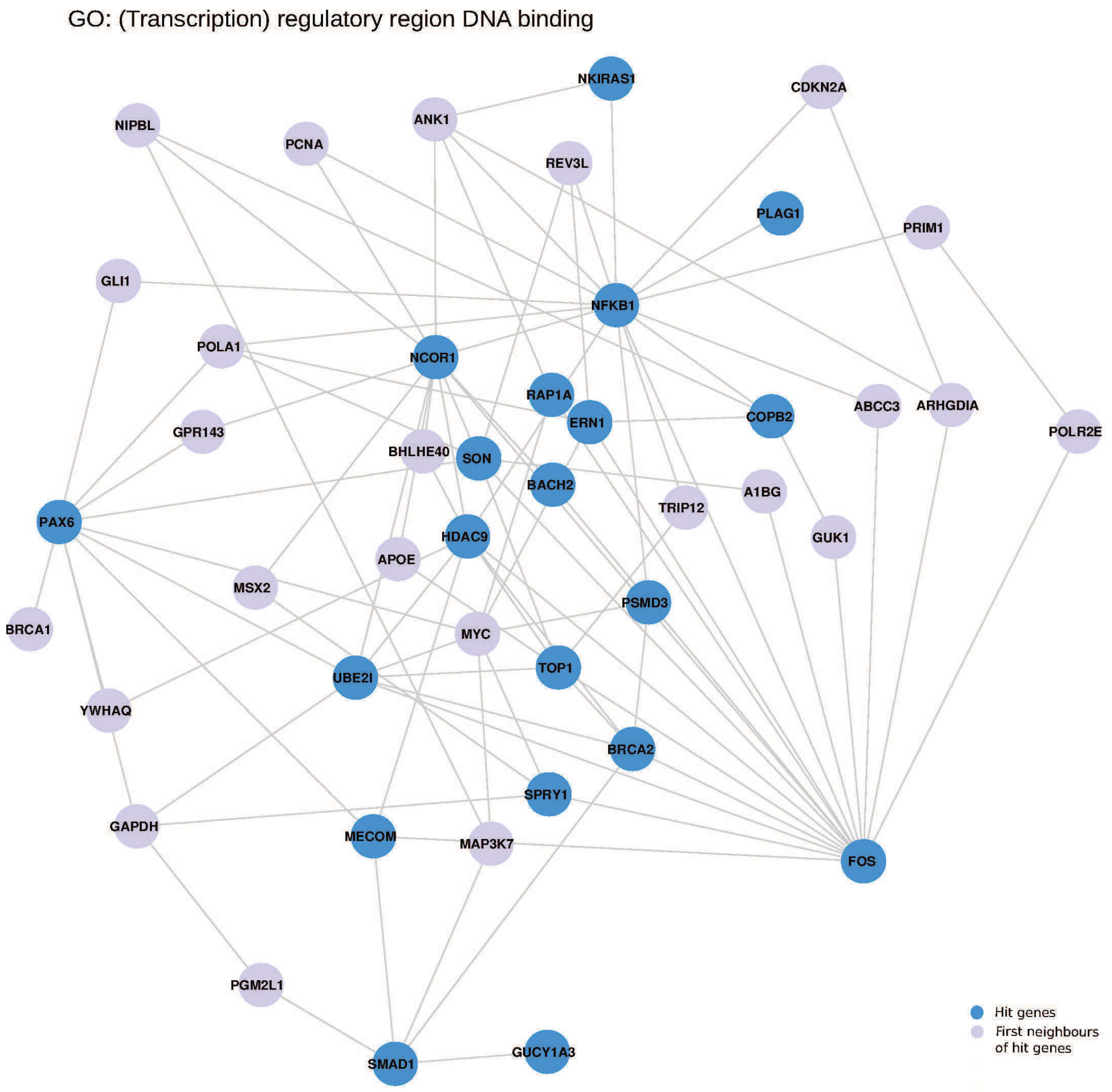


FIGURE 4.22: Blue nodes are the hit genes and purple nodes are their first neighbors.

20 2014. Their HGNC gene symbols were not directly available, thus the Entrez gene identifiers were used for WebGestalt GO enrichment. The multiple test adjustment was kept to BH and the minimum number of genes for a category was kept to three, for consistency with the previous network enrichment analysis. The significance level was altered though: instead of top 10 which was used in the previous GO analysis, here it was set to 0.05 to enrich with larger confidence for a category. Only three genes were enriched in many GO categories: ERN1, GSG2 and TTK. Two of these categories scored below the significance threshold: negative regulation of cell cycle, and cell cycle arrest. Thus, these three genes are significant regulators of cell cycle. Their interactions are visualized in Cytoscape and presented in Figure 4.23.

TABLE 4.11: Enriched GO terms with respective p-values

GO term	N^o **of genes enriched**	**Adjusted p-value**
negative regulation of biological process	19	6.1e-03
positive regulation of metabolic process	20	6.1e-03
macromolecule modification	17	6.1e-03
positive regulation of macromolecule metabolic process	14	6.1e-03
positive regulation of cellular metabolic process	14	7.3e-03
positive regulation of metabolic process	15	6.1e-03
positive regulation of cellular process	19	6.1e-03
cellular protein metabolic process	19	7.3e-03
protein modification process	16	7.3e-03
cellular protein modification process	16	7.3e-03
nucleic acid binding transcription factor activity	8	1.17e-02
chromatin binding	5	7.3e-03
sequence specific DNA-binding transcription factor activity	8	1.17e-02
SMAD binding	3	9.3e-03
enzyme binding	8	1.45e-02
DNA binding	13	1.6e-02
regulatory region nucleic acid binding	5	9.3e-03
regulatory region DNA binding	5	9.3e-03
sequence-specific DNA binding	7	9.3e-03
transcription regulatory DNA binding	5	9.3e-03
intracellular	38	1.4e-03
intracellular part	37	2.4e-03
organelle lumen	17	2.5e-03
intracellular organelle lumen	17	2.5e-03
nucleus	25	1.8e-03
nuclear part	19	4e-04
nucleoplasm	13	4e-04
nuclear lumen	17	8e-04
nuclear chromosome part	5	1.8e-03
nucleoplasm part	8	2.5e-03

4.12.0.4 Discussion on network enrichment

The Protein Interaction Network Module also revealed interesting modules, however almost covered by the two previous analyses. Many of the hit genes hold important roles in the human gene network and they are inter-connected. Surprisingly, almost all the genes (except CTNNA1 and EZH1) that we identified as KRAS SLPs are not directly connected with KRAS. So, they affect KRAS indirectly, and the signal passes through other nodes in its way. In conclusion, it could be supported that there are six genes that are significantly enriched for KRAS SLPs: i) the two direct interactors, CTNNA1 and EZH1, ii) the three enriched genes for cell cycle arrest: ERN1, GSG2 and TTK and iii) the very central and participating in almost all modules NFKB1 gene.

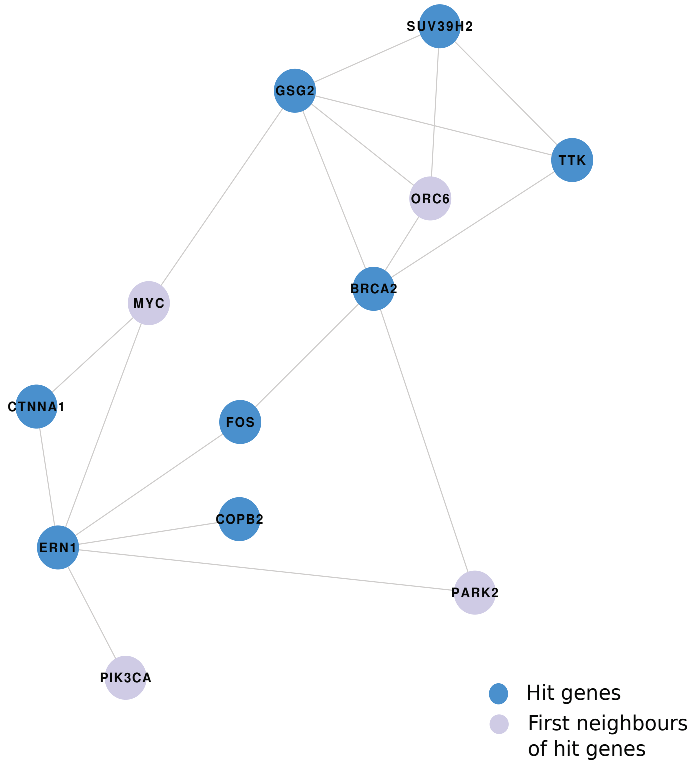


FIGURE 4.23: The connections of the three genes, ERN1, GSG2 and TTK, that are enriched when the human kinases are used as reference. Some other of the hit genes are present as well. Blue nodes are the hit genes and purple nodes are their first neighbors.

4.13 Final selection of genes for experimental validation

In this section, the results of all the previously mentioned methods are combined in Table 4.12. All the genes that were identified by at least two methods are considered and they are prioritized according to their total score (i.e. number of methods by which they were retrieved). A *hint paper*, as referred to in Table 4.12, is one that supports the respective gene as being very important in keeping the tumorigenic state of a KRAS mutant cell, without specifically mentioning that it is a KRAS SLP candidate (GSG). Here, the three genes for which there is such evidence, along with the respective source, are given:

1) ERN1: Evidence for this gene comes from the paper that introduced the *Barbie* screen [24]. It is found in the top 218 genes that, according to the authors, are more probable to be KRAS SLP candidates.

2) FOS: This is a gene downstream of KRAS and is one of the important factors in retaining the cancerous state of KRAS-dependent cells [98]. As member of the MAPK/ERK pathway it may be involved in the regulation of KRAS expression and for retaining the KRAS-induced tumors [99]

3) SPRY1: This Sprouty homolog 1 gene, is declared as antagonist of FGF signaling in Drosophila cells. It has been found that in KRAS mut cells it supports and facilitates EGFR signaling [100]. It thus acts as a positive feedback loop that retains the tumorigenic KRAS signaling.

After careful inspection of the table, the few top genes were concentrated, and provided to experimentalists, who were convinced to proceed with wet lab validation. The hit genes are the ones that have evidence from five or more sources in their favor: BRCA2, COPB2, ERN1, FOS, NFKB1 and SPRY1. Recall at this point that SPRY1 is the gene that was retrieved by all of the three basic methods (section 4.5). This is the first gene in which our experimental collaborators were interested. Moreover, quite a number of genes collected favorable evidence from four sources. These are CTNNA1, EZH1, HDAC9, NCOR1, PAX6, TOP1 and TTK. It has to be noticed that, although SMAD1 has a total score of 3, it is a GSG, a criteria that has more weight than the rest approaches. Thus, it can be considered a strong candidate.

Table 4.12: Sources of evidence for each gene

Gene/Method	Standard [a]	RSA	RIGER	RankAggreg	RNAiCut	Signif Phyper	GSG	Drug Screens	Hint paper	Network [b]	**Total**
BACH2	√				√	√					3
BCL2L13					√	√					2
BRCA2		√	√		√	√		√			**5**
COPB2	√		√	√	√	√					**5**
CTNNA1			√		√	√				√	4
DNTTIP1	√				√	√					3
ERN1	√				√	√		√	√ (Barbie)	√	**6**
EZH1			√		√	√				√	4
FOS	√		√		√	√		√	√		**6**
GSG2			√		√	√				√	4
HDAC9	√				√	√		√			4
ISL2	√				√						2
MECOM			√			√					2
MYBL1	√				√	√					3
NCOR1		√		√		√		√			4
NFKB1			√		√	√	√	√		√	**6**
PAX6			√		√	√		√			4
PLAG1	√				√						2
PSMD12		√	√								2
RAB7L1	√				√						2
RAP1A	√				√	√					3
SMAD1			√			√	√				3
SON			√		√	√					3
SPRY1	√	√	√	√	√				√		**6**
SUV39H2	√				√	√					3
TOP1		√			√	√		√			4
TTK	√				√	√				√	4
WASF3	√				√						2

---

[a] Intersection of hits among the three screens

[b] √: enriched for KRAS SLPs

4.14 Analysis and result on the incorporation of an extra or random screen

In section 3.5, a novel methodology, which examines if the inclusion of an extra or random screen alters the result, is described. Its results are presented hereafter. For each gene, the combination of screens in which its rank is minimized was found. Then, RankSLP examined which genes are assigned with their minimum ranking by the combination of Luo-Steckel and Wang screens and compared them with the original findings from these three screens' analysis. The encouraging observation was that, both when a random screen or when the Barbie screen are included, many of the detected hit genes are conserved.

More specifically, when the Barbie or a random screen was included, 15 and 13 genes respectively received their minimum *p*-value for the combination of the Luo, Steckel and Wang screens. The detailed findings are shown in Table 4.13.

Table 4.13: Genes enriched for best ranking by Luo-Steckel-Wang combination and respective *p*-values

Enriched gene when Barbie is included	**p-value**	**Enriched gene when random is included**	**Enriched gene and p-value**
BACH2	0.0023606279	CBL	0.0773371394
BRCA2	0.0548169947	COPB2	0.0130130515
CBL	0.0773371394	DNTTIP1	0.0086737373
COPB2	0.0130130515	FOS	0.0002950633
DNTTIP1	0.0086737373	HDAC9	0.0103203843
FOS	0.0002950633	ISL2	0.0290333177
HDAC9	0.0103203843	PLAG1	0.0011660877
ISL2	0.0290333177	RAB7L1	0.0257575380
PLAG1	0.0011660877	RRM1	0.0517120796
RAB7L1	0.0257575380	SPRY1	0.0294268330
RRM1	0.0517120796	SUV39H2	0.0021080075
SPRY1	0.0294268330	TRIB3	0.0345644340
SUV39H2	0.0021080075	UBE2I	0.0588745334
TRIB3	0.0345644340		
UBE2I	0.0588745334		

Among the genes that are enriched when the Barbie or a random screen is included, BRCA2, COPB2, FOS and SPRY1 have evidence from five or six sources according to RankSLP's analysis. HDAC9 has evidence from four sources. The genes BACH2, DNTTIP1, PLAG1, ISL2, RAB7L1 and SUV39H2, also appear in Table 4.12 but with less sources of evidence. Finally, UBE2I is not seen for the first time; it was detected by Standard method when the best ranking shRNA was considered. However, it was not confirmed by any of the other methods, that is why it is not reported in Table

4.12. The only genes that are new are CBL, TRIB3 and RRM1. Among these, CBL is very strongly connected with a network to SPRY1, one of RankSLP's most potential candidates. Moreover, the p-values of these additional genes are not very significant (except TRIB3). In conclusion, according to the random or extra screen analysis, the majority of the genes identified by the methodology implemented in this thesis are confirmed robust candidates.

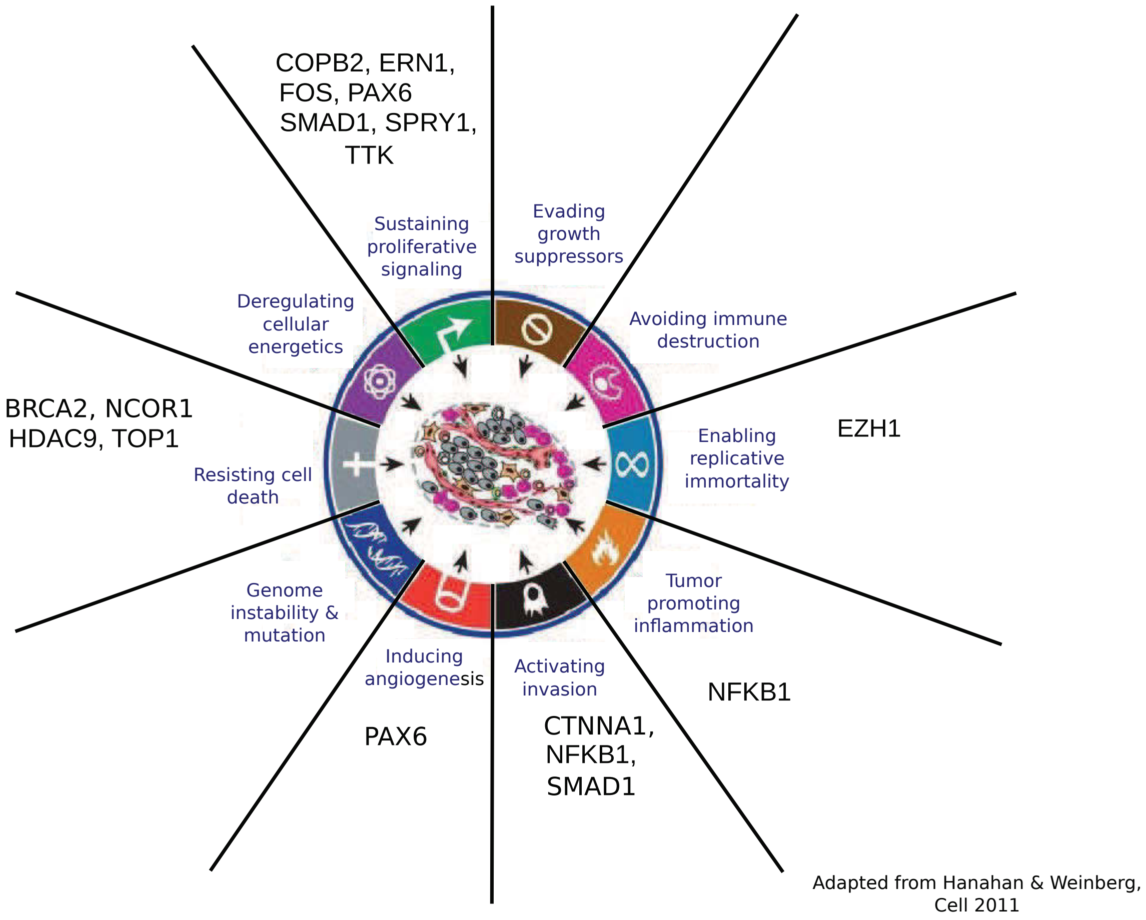
Chapter 5

Discussion

This Chapter further discusses the findings of this thesis. First of all, the selected most probable KRAS SLP candidates are deeper investigated and associated with the hallmarks of cancer. These are the properties of cancer cells, as proposed by Hanahan and Weinberg [101]. Moreover, the relatively small overlap between the genes that are detected by this analysis and the *Gold Standard Genes* is discussed. This Chapter also provides an evaluation of the three basic ranking methods that are applied in this work: Standard method, RIGER and RSA. In addition, the Cross Entropy rank aggregation is compared with baseline rank aggregation techniques. Furthermore, the methods and the parameter selection decisions are assessed here. Finally, this Chapter discusses the contribution of this thesis and compares it with existing work in the field.

5.1 Final selected genes and their roles in cancer development

Hanahan and Weinberg, in their seminal paper in 2000 [102], proposed the six properties that are shared among cancer cells, and showed that they are necessary for tumour initiation and expansion. They call these properties *"hallmarks of cancer"*. They further expanded this initial set of six hallmarks to nine in their 2011 paper [101]. A cancer cell usually exhibits the characteristics of more than one hallmark category. These hallmarks are widely established in cancer biology and are often used as a reference. The functionalities of the candidate KRAS SLPs with evidence from four or more sources (see Table 4.12) are explored and associated with the nine, updated hallmarks. To this aim, PudMed-Gene[1] and Gene Cards[2] were interrogated (access date: 23 July 2014), and the following list was populated.


COPB2, ERN1,
FOS, PAX6
SMAD1, SPRY1,
TTK
Sustaining proliferative signaling
Evading growth suppressors
Deregulating cellular energetics
Avoiding immune destruction
BRCA2, NCOR1
HDAC9, TOP1
Enabling replicative immortality
EZH1
Resisting cell death
Genome instability & mutation
Tumor promoting inflammation
Inducing angiogenesis
Activating invasion
NFKB1
PAX6
CTNNA1,
NFKB1,
SMAD1

Adapted from Hanahan & Weinberg, Cell 2011

Figure 5.1: This figure it adapted from the latest paper of Hanahan and Weinberg [101]. Alongside the cycle, the nine hallmarks of cancer are presented with the use of a characteristic symbol for each of them. The rays of the cycle are manually expanded to surround the hit genes that are associated with the respective hallmark.

BRCA2: It is involved in maintenance of genome stability, as it is responsible for the repair of double strand breaks. BRCA2 is considered a tumor suppressor gene, as tumors with BRCA2 mutations generally exhibit loss of heterozygosity (LOH) of the wild-type allele. That is why it is associated with the *resisting cell death* hallmark.

CTNNA1: This gene participates in catenin-cadherin binding and is important in keeping cadherin cell-adhesion properties. It is reported as a potential crucial player in cell differentiation. The annotation of this gene is not very informative. I hypothesize that it participates in epithelial-to-mesenchymal transition, thus it is associated with *activating invasion.*

COPB2: This gene is a member of the Golgi coatomer complex and is essential for

---

[1] http://www.ncbi.nlm.nih.gov/gene/

[2] http://www.genecards.org/

Golgi budding and vesicular trafficking. Due to its role in signal transduction, it is reasonably associated with *sustaining proliferative signaling.*

ERN1: ERN1 is important in altering gene expression as a response to endoplasmic reticulum-based stress signals, thus it is assigned to *sustaining proliferative signaling.*

EZH1: EZH1 interferes with methylation of histone H3 and helps in maintaining embryonic stem cell pluripotency and plasticity. It is required for embryonic stem cell derivation and self-renewal. In the Hanahan-Weinbeg cycle, it is associated with *enabling replicative immortality.*

FOS: This gene is widely known as regulator of cell proliferation, differentiation, and transformation and sometimes as responsible for apoptotic cell death. It is also a member of the MEK/ERK pathway and may have a positive feedback role in retaining KRAS tumorigenic state. Hence it is placed in the *sustaining proliferative signaling* category.

HDAC9: This is a histone deacetylation protein and affects the way transcription factors regulate DNA transcription by altering chromosome structure. Depending on what genes are covered by the histone that is deacetylated, this gene can have various effects on the cell's life. Here it is put in the *resisting cell death* category, but it could be assigned in other parts of the cycle as well.

NCOR1: This gene negatively controls transcriptional repression of thyroid hormone and retinoic-acid receptors by sponsoring chromatin condensation. It is part of a complex which also includes histone deacetylases and transcriptional regulators. This is confirmed by the network analysis, where it is found to be directly connected to HDAC9 (section 4.12). It is thus put in the same hallmark with HDAC: *resisting cell death.*

NFKB1: NFKB1 is a very influential transcription regulator. Its inappropriate activation leads to diseases of the inflammatory system and its inhibition to problematic development of the cells's immune system. It is thus playing an important role in the cell's inflammation and invasion (immune system cells are particularly invasive). Therefore it is assigned to two hallmark categories: *activating invasion* and *tumor promoting inflammation.*

PAX6: This gene contains DNA-binding domains which regulate gene transcription, thus it is put in the *sustaining proliferative signaling* category. It is also required in the differentiation of pancreatic islet alpha cells. As found in the STRING DB[3], accessed on 23 July 2014, it is connected to gene IPO13. The latter mediates the import of specific cargo proteins from the cytoplasm to the nucleus and is dependent on the Ras-related nuclear protein-GTPase system. PAX6 is furthermore connected to the transcription factors SOX2 and SOX3, which participate and influence development of the embryo and cell fate. It is also found to promote blood vessel creation.

SMAD1: SMAD1 is a member of the SMAD family of proteins, which possesses signal transducing and transcriptional modulating capacities. It affects multiple cellular

---

[3] `http://string-db.org/`

pathways and influences development, immune system response and other crucial cell functions. It is also critical for blocking PAI, so it possibly regulates the invasion of tumour cells. Thus, it is connected with two hallmark categories: *sustaining proliferative signaling* and *activating invasion*.

SPRY1: The Sprouty homolog 1 gene, is declared as antagonist of FGF signaling in Drosophila cells. In normal cells it acts as an inhibitor of FGF and EGF signaling pathway activation, since it negatively regulates Receptor Tyrosine Kinases (RTKs). However, in KRAS mutant cells, it has been found to act in the opposite fashion and actually support and facilitate EGFR signaling [100]. It is clearly connected to the *sustaining proliferative signaling* hallmark.

TOP1: This gene encodes a DNA topoisomerase, an enzyme that controls and alters the topologic states of DNA during transcription. This enzyme catalyzes the transient breaking and rejoining of a single strand of DNA, which allows the strands to pass through one another, thus altering the topology of DNA. Due to its property to counterbalance DNA damage, it is placed to the *resisting cell death* hallmark.

TTK: This gene encodes a protein kinase which is able to phosphorylate tyrosine, serine and threonine. It is essential for chromosome alignment at the centromere during mitosis and for centrosome duplication. It is thus crucial for cell proliferation, and therefore assigned to *sustaining proliferative signaling*.

The majority of the hit genes are predominantly associated with the *sustaining proliferative signaling* hallmark. This corroborates Hanahan and Weinberg's claim: "Arguably the most fundamental trait of cancer cells involves their ability to sustain chronic proliferation". They support that this hallmark is the most commonly represented in cancer cells, since it clearly is an accelerator of tumour development. It is consequently stronger compared to tumor suppressor genes. Of course, as they claim, "often these signals influence yet other cell-biological properties, such as cell survival and energy metabolism" [101].

Most hit genes can be associated with at least one hallmark. This confirmation ascertains that the scientific methodology followed in this work is rational and in accordance with established biological knowledge. However, the analysis of the identified genes may be expanded as additional roles are identified. The specific subdivision on Hanahan and Weinberg hallmarks presents the authors opinion.

5.2 Overlap of Gold Standard Genes with RankSLP findings

As shown in section 4.10, there is a moderate overlap between the genes retrieved in this thesis and the GSGs reported in literature. Please note that with GSGs here, the overlap of all GSGs with the genes that are commonly screened by the three main screens (Luo, Steckel and Wang) is meant. Only two of the GSGs, NFKB1 and SMAD1, are identified as hits by the overall analysis, a fact that nominates them of being stronger candidates. The rest GSGs achieve moderate to high rankings, but not the highest. This finding is positive since it is in accordance with our hypothesis; existing studies from which the GSGs were extracted mainly concentrate on the top genes. Another explanation for that could be the inherent to RNAi screening OTEs, extensively presented in section 2.2.3. Some of the GSGs may in reality be screen artifacts and this implies that one shouldn't rely too much on them. In the current context, however, there exists nothing else to help evaluate the findings. Therefore the GSGs were considered for validation, having awareness of their limited evaluation power.

Despite these problems with GSGs, the hypothesis that a gene may not rank at the top in any of the screens but relatively high in all of them revealed new potent candidate KRAS SLPs, was confirmed by a multitude of methods. The confirmation of two of the existing GSGs by RankSLP is, on the other hand, indicating that its methodology is rational and in accordance with existing biological knowledge.

5.3 Overview of methods used in this thesis

In this section it is adressedd why the specific methods were chosen by RankSLP. Are there additional methods that could be applied and how would they alter the findings?

5.3.1 Three basic methods for hit detection

At first, the three main methods that RankSLP applied are examined. They are "main" because they perform the first, crucial filtering for hit genes.

- Standard Method: This method is essentially retrieving the hit genes by filtering them using two types of criteria: i) Mean ± k Standard Deviations or Median ± k MAD and ii) p value < a defined threshold. Both criteria have been extensively used for detection of hits from RNAi screens [103], [104], [105], [25].

- RIGER: The RIGER method, provided by BROAD Institute of MIT and Harvard, takes into account the background distribution, the difference between the mutant and wild type classes and the gene enrichment. Therefore, it satisfies the required criteria for method selection which are set at section 3.2. RIGER is used for hit detection in the case of pooled shRNA or siRNA screens and has been applied by Luo *et.* al. [25] and Barbie *et.* al. [24]. These publications and their respective datasets constitute a basis for RankSLP's analysis, and the methods they suggest were highly considerable. According to the original work where RIGER was introduced [54], RIGER has very good performance on real datasets.
- RSA: In accordance to the RIGER method, RSA method takes into account the background distribution and the difference between the mutant and wild type classes. According to König *et* al. [56], it performs astonishingly better than the activity-based method as its hits have higher reconfirmation rates on follow-up screens. Birmingham *et* al. in their review paper [105], suggest RSA to be the most robust method for RNAi screening hit detection, when multiple shRNAs target the same gene. It is also incorporated in the stability rankings algorithm [106] and reported to score very well compared to the other alternatives of the algorithm.

It has to be noticed that both RSA and RIGER are easily implemented, easily interpreted and most widely used. They are furthermore applied as combinatorial methods to the Standard method by other studies in the field of integration [2], [107].

There are additional methods that could be used to identify hit genes from RNAi screens, according to the review [105]. These are

- **S**trictly **S**tandardized **M**ean **D**ifference (SSMD): This method is analogous to the standard z-score when no replicate samples exist. In the case of replicate samples however, SSMD allows for control of both false positive and false negative rates. In addition, it is not dependent on sample size. These two are advantages that increase its reliability. It has been recently implemented in [108].
- **R**ank **P**roducts (RP): This method was originally developed for hit identification from microarray studies [109]. In the case of replicate RNAi screens, its premise is that a sample which ranks high in one sample should rank high in all samples.
- **S**ignificance **A**nalysis of **M**icroarrays: This method was developed by Tusher *et.* al in 2001 [110]. This methods assigns a score to each gene by measuring the gene's expression and correcting it based on the standard deviation of repeated

measurements. Essentially, it is another robust way of detecting hit differentially expressed genes between two classes.

RP and SAM methods are in the same direction with t-test that has already been used by RankSLP. Given that, at the end, multiple methods are combined and their results are compared with external data, it was decided that RP and SAM application would not provide an additive advantage to the current work. SSMD on the other hand is very robust and different from the methods used so far in this analysis. It has not been used in this work because it requires extensive screen plate information and image data which were not available to us. The incorporation of SSMD analysis could be a very potent future addition in this work, once the respective data are retrieved and explained by the providers. However, RIGER and RSA are very robust methods as well towards the detection of hit genes from replicate RNAi screens.

5.3.2 Evaluation of Standard, RSA and RIGER methods: Which performs best?

This part discusses the output of the three basic methods that were applied in the current analysis. An effort to evaluate their performance based on how many of the final selected genes they retrieve is made (see section 4.13). The set of final retrieved high-confidence genes consists of BRCA2, COPB2, ERN1, FOS, NFKB1, SMAD1 and SPRY1.

- **Standard method:** It retrieved COPB2, ERN1, FOS and SPRY1 out of its 20 totally retrieved genes.
- **RSA:** It retrieved BRCA2, SPRY1 out its 9 hit genes.
- **RIGER:** It retrieved BRCA2, COPB2, FOS, NFKB1 and SPRY1. In total RIGER retrieved 17 genes.

The Standard method pinpoints four final hits among its 20 genes, which is translated into 20% retrieval capacity. RSA, on the other hand, retrieves only two of the strongest candidates which means $\sim 22\%$ recall. This method is very specific and, although the retrieval capacity is similar to that of the standard method, in absolute quantities only two of the strong candidates are captured, a fact that does render this method less appropriate if one wanted to use only one method towards hit gene detection. Finally, the RIGER method detects six out of seven strongest candidates. Having a total retrieval of 17 genes, this number translates to $\sim 29\%$ recall. In addition, the absolute numbers of hit gene recall are higher in RIGER.

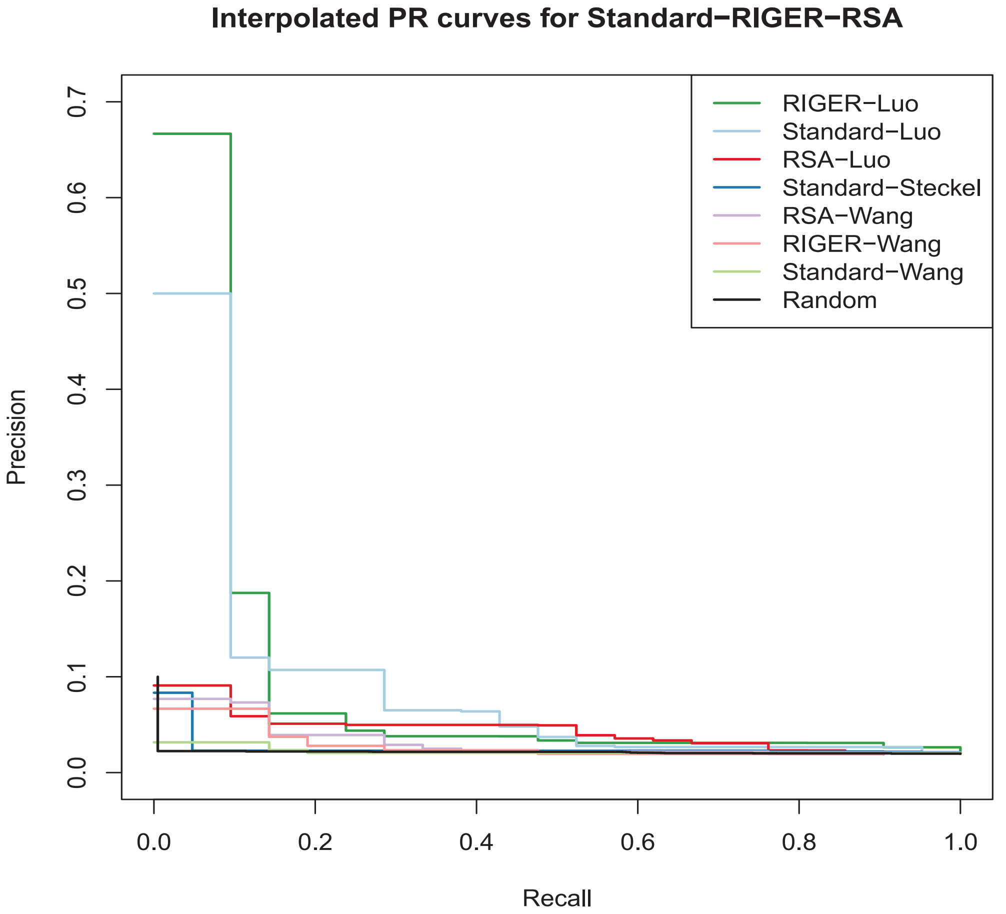


FIGURE 5.2: The overall precision is not very large. This was expected, due to the low evaluation power of the GSGs (explained in section 5.2). However, an encouraging finding is that the largest precision values are met in the top of the ranked lists, which is the aim of this thesis. Another positive result is that all methods outperform the random ranking. In comparative terms, best precision for the top of the ranked lists is achieved by the application of the RIGER method on the Luo screen. Second best is the Standard method for Luo. Standard method for Steckel and RSA for Wang and for Luo follow.

Figure 5.2 depicts the precision-recall curves for all the Standard, the RIGER and the RSA methods, along with a random ranking (in black). All the applied techniques outperform the random case.

In Figure 5.2 it can be seen that all three methods contribute to the final set of genes. This is also explained in the previous Chapters (Data and Methods, Results); the final RankSLP's result stems from all methods' combination. However, if one had to apply a single method, this would be RIGER, as it detects most of the genes that are further subject to experimental validation. Moreover, RIGER achieves the largest overall precision for the top of the Luo ranked list. It seems that the RIGER approach of calculating the signal to noise ratio and the weighted sum of the shRNAs that target a gene, followed by gene enrichment analysis and normalization, is suitable for RNAi screen analysis. These are encouraging news for biologists, given that the RIGER method tool can be freely downloaded from its main page[4] and executed through a friendly Graphical User Interface (GUI), which can be used by someone without any computational background.

5.3.3 Rank aggregation methods

RankSLP uses the RankAggreg() [59] method from R since the whole pipeline is developed with this language. Pihur *et* al. [59] show that their function works very well on real data. In this subsection, a comparison between the RankAggreg() procedure and baseline rank aggregation methods is performed.

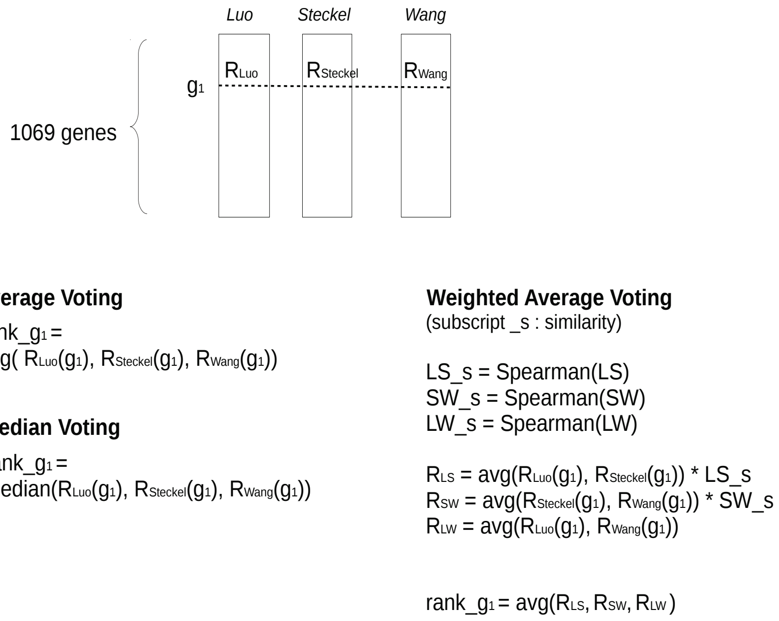

Luo
Steckel
Wang
1069 genes
g1
RLuo
RSteckel
RWang
Average Voting
rank_g1 =
avg(RLuo(g1), RSteckel(g1), RWang(g1))
Median Voting
rank_g1 =
median(RLuo(g1), RSteckel(g1), RWang(g1))
Weighted Average Voting
(subscript _s : similarity)
LS_s = Spearman(LS)
SW_s = Spearman(SW)
LW_s = Spearman(LW)
RLS = avg(RLuo(g1), RSteckel(g1)) * LS_s
RSW = avg(RSteckel(g1), RWang(g1)) * SW_s
RLW = avg(RLuo(g1), RWang(g1))
rank_g1 = avg(RLS, RSW, RLW)

Figure 5.3: The comparison is made with respect to three baseline methods: Average voting, median voting and weighted average voting. Luo, Steckel and Wang are the three ranked input lists that correspond to each screen's ranking by Standard method. The best ranking shRNA is considered when many shRNAs target the same gene (Luo and Wang case). R_{Luo}, $R_{Steckel}$ and R_{Wang} are the rankings of the common gene g_1 as calculated by each of the screens. $rank\_g_1$ is the ranking of gene g_1 as calculated by the three baseline methods.

The respective Precision-Recall curves which evaluate the performance of the rank aggregation methods are shown in Figure 5.4. The comparison is made on the top 10%

---

[4] `http://www.broadinstitute.org/cancer/software/GENE-E/`

retrieved features by all the methods because the RankAggreg() procedure was implemented only for this part of the ranking lists, due to complexity reasons (see subsection 3.4.1).

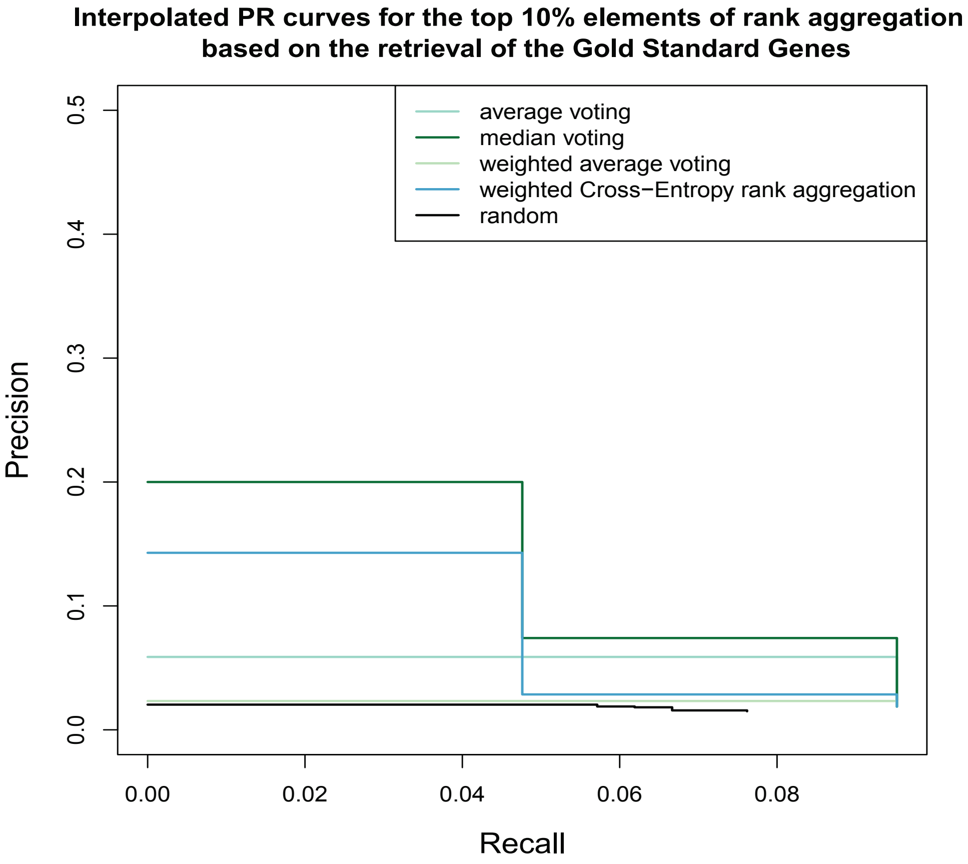


Figure 5.4: Evaluation of the top 107 (top 10%) genes retrieved by different rank aggregation methods based on the retrieval of the 21 common in all screens GSGs. All the methods retrieve two of the GSGs in their top 10% but in different ranking positions. The largest precision is achieved by median voting, followed by the weighted Cross Entropy rank aggregation that was implemented in RankSLP.

They all outperform the random case and all apart from the weighted average voting retrieve two out of the 21 GSGs in their top 10%. When the remaining collected genes are investigated, the average, median and weighted average voting approaches select 23, 24 and 20 genes respectively, intersecting with the hit 40 genes initially selected by Standard, RIGER and RSA methods (section 4.5). RankAggreg, on the other hand, retrieves only three genes in its top 107 ranks: COPB2, NCOR1 and SPRY1. This does not necessarily mean that it is worse (as PR curve shows that it performs similarly with the rest) but that it is more specific and targeted. It was applied in this work because it implements entropy, which is a benchmark and extensively used approach for the solution of optimization problems. A noticeable point is the good performance of the median rank aggregation: This corroborates the proof by Dwork *et* al. that the median consensus ranking is a good aproximation of the Spearman distance criteria between the final *super*-list and the input rankings [51].

5.4 Assessment of screen quality

From the previous section, the Steckel screen seems of lowest quality; The standard method on Steckel screen performs worse than RIGER and than Standard method on Wang and Luo screens. Moreover, the best performing RIGER method is actually applied on two screens only: Luo and Wang. The Steckel screen is not taken at all into consideration as it is not performed with many shRNAs targeting the same gene. Since RIGER still retrieves more of the strongest candidates compared to the other two methods, the quality and relevance of the Steckel screen is interrogated. In addition to that, the weight that is attributed to Steckel screen before the execution of *RankAggreg()* (section 4.8) is 0.1, in opposition with Luo and Wang screen coefficients which are 0.6 and 0.3 respectively.

Considering these findings, one could argue that Steckel screen could be omitted if the aim is a handful of candidate SLPs and there are time limitations for final result retrieval. For the current thesis though, it is useful enough to observe that the rescued genes from the other two screens are also important in the Steckel screen. This supports the hypothesis that a relatively highly ranked gene in one screen can be also relatively highly ranked in another screen, which makes it a more probable candidate.

To quantitatively assess the quality of the Steckel dataset ranking, a metric often adopted in Information Retieval (IR) and more specifically in Natural Language Processing (NLP) approaches was employed. This is the Normalized Discounted Cumulative Gain (NDCG). NDCG measures the relevance of a feature based on its position in the retrieved result list, using a graded relevance scale of features in a search engine result set. It imposes a penalty on the features (here genes) that are retrieved lower than they should in the ranked list. It applies a logarithmic reduction factor for smoother discount of these cases. The formula of DCG is given by

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i},$$

where rel_i is the graded relevance result at position i and p is a particular rank position. Similar to precision at k, it is evaluated over some number k of top search findings. The normalized measure NDCG is given by dividing the DCG with the maximum possible achieved DCG at position p. This is a real number between 0 and 1; the closer it is to 1, the best the information retrieval capacity of the applied method.

In the present case, the ranked Steckel gene list, as calculated by the Standard method, was the retrieved result list of which the relevance was calculated. The ideal ranking with which it was compared, is a ranked list in which all GSGs are ranked in the top positions. The NDCG result for the applied threshold on the Steckel ranked list ($z < -1$) was ~ 0.28. For comparison, the NDCG metric for the Luo and Wang rankings as retrieved by the Standard method was calculated. These are ~ 0.55 and ~ 0.28 respectively. This result shows that the Steckel screen is not of that bad quality and probably if more values per gene were provided in this dataset, then the RIGER method could be applied on it with expected good results.

Moreover, the NDCG metric shows that the Luo screen ranking is of best quality. Generally, Luo screen is more reliable as also proved from Figure 5.2, where the Luo screen achieves the highest precision at the 10%-20% recall level.

5.5 How does the selection of thresholds and methods influence the final SLPs?

In this section, a critical view of the methodologies and the threshold criteria that were applied in this work are discussed. Did the decisions that RankSLP took in various cases influence its final results? Some of its choices are very deeply investigated but some others allow room for improvement. Critical steps of the analysis and assessment of their stability are provided in table 5.1.

Table 5.1: Assessment of methods and thresholds chosen

Good Assessment	**Needs improvement**
Cell line variability	The amount of genes tested
RankAggreg() top 10%	Threshold selection for the main methods
RankAggreg() distance measure	Best ranking shRNA used in some functions
Statistical test	WebGestalt thresholds
RNAiCut threshold	STITCH-drug datasets compound matching
STITCH drug-target confidence threshold	Drug to target assignment
Venn diagrams & *phyper*	
Comparison with GSGs	
Mapping of gene identifiers	

In what follows, table 5.1 is explained, starting with the *Good Assessment* column and proceeding from top to bottom.

Cell line variability can in theory influence the findings. Taking that into account, RankSLP carefully selected the datasets so that they consist of isogenic cell lines, differing only in the KRAS mutation, a fact that minimizes variability.

For the *RankAggreg()* R function, RankSLP searched for the most significant features in the top 10% and 5% of genes from the three screens. That is because the *RankAggreg()* method is of high complexity and the optimal solution is guaranteed only in the case that the input lists do not contain more than around 100 elements. As far as the data used in this work are concerned, the top 10% elements for each list correspond to 107 genes. Moreover, the distance measure (Spearman's footrule distance) and the other parameters of *RankAggreg()* are carefully selected and this selection is justified in 3.4.1 and in 4.8

The statistical test that was applied in the case of Luo and Wang screens was the paired t-test, because the samples are paired (isogenic cell lines and three replicates corresponding to the same well of the 96-well-plate) and only one independent variable, the viability of the cell line, is measured. The distribution can be considered normal with only three samples.

As far as the RNAiCut method is concerned, for each set of top-k samples, an underlying PPI network of a size at least as the size of k is sought. The returned network is the one for which the first local minimum p-value is met. In the Luo case, this first local minimum was too small; Thus, after correspondence with the developers of the method, the next one was selected.

The drug-target matching confidence threshold was set to 70% according to the suggestion of STITCH creators. This choice is explained in subsection 4.11.4 and shown in Figure 4.18.

Regarding the Luo and Wang screens, where multiple shRNAs target the same gene, more than one shRNAs for the same gene often pass the threshold criteria. As shown in section 4.5, the intersections of genes of which multiple shRNAs rank below threshold are evaluated using a hypergeometric distribution test. This test returned a clear p-value and RankSLP simply selected the genes in the significant intersections (p-value < 0.05).

Furthermore, existing literature and past findings have extensively and in detail been investigated. Hence, the list of GSGs with which the comparison is done is full.

The mapping of identifiers among the screens and the conclusion in 1069 common genes is also performed very carefully, by conversion to HGNC identifiers. In some cases, the same gene is knocked down among the screens but it is reported by a commonly used symbol in all three of them, which is not the HGNC identifier. These cases are also investigated in detail.

On the other hand, the second column of table 5.1 contains some of the RankSLP's decisions which allow improvement. The final set of tested genes for synthetic lethality with KRAS, are the common genes among all screens. This means that if a gene is ranking very high in one screen but is not even screened in the second, it will not even be investigated. On the one side, the datasets with fewer genes (Wang, Steckel) contain a pre-selection of cancer related proteins, thus more probable KRAS SLPs. On the other side, a not-so well annotated gene may be a strong candidate. The following elaboration could be considered in the future: After an initial examination of the RNAi screens based on the common genes, one should proceed with the drug-screen analysis part. If a gene is retrieved as hit from the drug screens but is not screened in all the RNAi screens, its performance should be investigated in as many screens as it is present.

A question that was risen, is if the thresholds that RankSLP selected were arbitrary and how their choice may affect the findings. Before proceeding to the p-value threshold of 0.05, RankSLP had experimented with the Standard, RIGER and RSA methods and with cutting at different thresholds from the top of the ranked list: 10%, 20% and 30%. It then applied the 0.05 p-value criterion, as this was the prevailing thresholds used by RIGER and RSA methods. The significance threshold for the Standard method was set to p-value < 0.05 for consistency with RIGER and RSA. For the Steckel screen, where no replicates are provided, the relaxed z-score < -1 criteria was applied. The genes that intersected among the screens and among the methods were not very much different from the ones that were retrieved by the higher thresholds of 10%, 20% or 30%. These first findings were not restrictive. As shown in the Results section, seven other ways to confirm these genes were incorporated for the final selection of the KRAS SLP candidates. Thus, the thresholds were not so influential. Overall, each method provided some highly potent genes but at the end, all the methods were combined and the genes that were *relatively high* among all are reported. The final resulting genes are quite robust to these thresholds because they are rescued after integration of multiple methods. However, there is room for further tuning in selection of thresholds.

In the cases where only one out of multiple shRNAs that target the same gene should be chosen from the Standard method findings, the best ranking shRNA was selected. This is actually the shRNA with the strongest knock-down effect. RankSLP used this approach because of the reasoning that if an shRNA depletes a gene, then the resulting phenotype should be gene knock down, as this strong effect will be more dramatic for the gene. In addition, experts in the field usually consider the shRNA with the strongest effect against a gene of interest, given that there is at least another shRNA targeting the same gene, which leads to a similar viability value [24, 25, 111]. That is why the existence of more than one shRNAs against a gene that rank below threshold was also considered by RankSLP (see section 4.6). However, this is not a trivial problem and a more sophisticated method could be developed in the future.

The criteria and thresholds used for the WebGestalt analysis are the default ones. They were consistent for the different runs of the algorithm that were conducted but more alternatives could be investigated in the future.

Something that requires improvement, is the mapping of drugs between STITCH and the drug datasets. STITCH attaches to each drug its corresponding inchi key, which is a unique identifier. On the contrary, as mentioned in 4.11.4, the compound inhibition screens provide only the aliases for each compound, making very difficult the retrieval of common drugs between each screen and STITCH.

Finally, in the comparisons with high throughput drug screens conducted in this work (see section 4.11), the assumption made is that as long as a specific protein is among a drug's targets it is knocked down by it. This is an simplified approach as usually a drug targets more than one genes which may interact with each other; the drug's effect on the one gene may counter-act the drug's effect on the other gene and many, non -linear relationships may evolve. However, there is evidence that in some cases this oversimplified binary assumption worked on drug-target interaction assessment [112, 113]. For RankSLP the evidence that a gene is targeted by a drug with differential effect between the KRAS mutant and the wild type cells suffices for inclusion of the target gene in the SLP candidates.

5.6 Connection of this work with previous methods and findings

In this thesis, a novel method which takes as input various synthetic-lethal for a specific gene RNAi screens and analyzes them integratively was developed. The result is a set of potential SLPs of the gene, detected by a purely in vitro methodology. The approach is general and can be applied to other oncogenes, like MYC. MYC is another currently undruggable oncogene, the SLPs of which are investigated. In the search of potential computational methods, developed towards targeting MYC SLPs, PubMed has been queried with *"MYC synthetic lethal"* as of 24 July 2014. There is not a single study that tries to identify candidate MYC SLPs by using purely computational methods. Some studies perform high throughput RNAi screens [114], [115], of which the findings are further analyzed using bioinformatics. Since MYC is the only oncogene apart from

KRAS which is so well known and investigated, it can be claimed that RankSLP is a novel method that can be used in the context of MYC as well. The existence of bioinformatic methods specific to the detection of synthetic lethal pairs has been further interrogated. A poster, where the boolean nature of synthetic lethal pairs is considered [116] has been detected. However this work is at a preliminary stage yet and no published results exist.

From the bological point of view, previous approaches towards the treatment of KRAS-dependent tumours suggested dual disruption of the PI3K/Akt/mTOR and the MEK/ERK pathways, especially in the case of pancreatic KRAS-dependent cancer [117, 118]. This is reasonable since KRAS in its mutant state is needed for both MEK/ERK and PI3K pathway activity, but not in its wild type. The first pre-clinical studies in Non Small Cell Lung Cancer (NSCLC) which followed this combinatorial approach had impressive results [119]. In 2013, Downward *et* al. followed a similar approach in NCLSC cells, this time by minimizing induced toxicity, having again encouraging results [120]. It is of future work to bring all these methodologies to clinics. One of the top genes retrieved by the current analysis is NFKB1. This gene participates in MAPK/PI3K pathways, a fact that is in accordance with the previous suggestion and exhibits high potential of positive clinical results.

Apart form NFKB1, RankSLP detected a set of genes, some of which were not reported in the past and could be worth looked further into.

Chapter 6

Conclusions and future extensions

This final Chapter summarizes the achievements of this thesis towards the solution of the open problems, presented in section 1.1. It emphasizes the complexity of the calculations and the contributions to biology and computer science. Furthermore, possible future extensions are presented.

6.1 Summary

In this work, a robust set of KRAS SLPs is sought, by investigating the molecular characteristics of the processes in which mutant KRAS participates and its mechanism of action. The detection of KRAS SLPs is important for the advancement of human health, because these can provide new targets for the treatment of KRAS-dependent cancers (i.e. pancreatic cancer). Many studies have been conducted towards the detection of KRAS SLPs, either RNAi or compound inhibition screens. A number of SLPs is detected by each of them but, as claimed in section 4.1.3, the results rarely agree. This variation can in part be explained by the use of genetically different cell lines, different methods used for quantifying cell viability, and technical and biological noise. However, this inconsistency implies that the selected genes by each study are probably able to work only on the specific dataset and not on new data. Sufficient additional and un-exploited information should be present in these existing high throughput sets; data that have been underestimated in the past can be analyzed collectively and reveal more stable KRAS SLPs, having legitimate results in more than one datasets.

In this thesis, a combination of computational approaches on data ranking and hit detection was applied towards the retrieval of KRAS SLPs, through integration of three existing high throughput KRAS-specific RNAi screens. Many considerations regarding this integration potential were risen. From an external point of view it may seem like trying to "compare apples and pears", due to the many differences between the data. However, formulation of some criteria and definition of rules that must be fulfilled so that dataset integration can be possible was achieved (see 3, sections 4.3, 4.4, 3.2). Although the application of such rules may restrict the search space, at least it ensures the robustness of the findings, since they are confirmed by more than one experimental datasets. In addition, a set of external data for evaluation of intermediate findings has been utilized. The role of the retrieved hit genes in cancer has been analyzed, with respect to the widely accepted Hanahan and Weinberg *"hallmarks of cancer"* [101]. These *hallmarks* represent nine properties of cancer cells to establish a solid tumor (see section 5.1). The detected hits were found to fit to one or more of these *hallmarks*. In conclusion, seven genes that are confirmed by the majority of applied approaches and thus concentrate highest evidence of being potent KRAS SLPs, are reported as hits. These are: BRCA2, COPB2, ERN1, FOS, NFKB1, SMAD1 and SPRY1. It is suggested that these genes can further be tested in vivo and, if successful, be brought to clinical trials. Yet, the possibility of detecting an additional potential candidate in their proximity or their influential network should not be ignored. The reason for that is exactly the limitation in genes that were screened and in the provided cell lines.

6.2 Complexity

The overall time that RankSLP needs for completion of its calculations is ∼18 CPU hours on one Core of an Intel X5650 CPU, 2.83Hz, 96 GB of memory. The datasets were transferred to a local partition to avoid the time spent fetching and writing them from and to the hard disk. It is considerable that the datasets downloaded from STITCH and used in the analysis occupy 23GB, since they consist of millions of drug-target interactions and drug identifiers. Most of the time is spent on the calculations on STITCH and on the drug datasets (11 CPU hours) and on the Cross Entropy Rank Aggregation (6 CPU hours). The respective procedures are explained in sections 4.11 and 2.3, 3.4. The rest of the RankSLP process takes around one CPU hour. These are the durations of the final version of RankSLP, after having tuned all the parameters and having concluded to the final variable values. A lot of time was spent on dataset selection and parameter optimization. The most crucial bottlenecks were faced at:

- FDR calculation: Initially, an effort was made to calculate the False Discovery Rate (FDR) of the three RNAi screens with replicate information (Luo, Wang, Barbie), aiming to further judge screen quality based on the shape of their empirical cumulative distribution function (ecdf). However, these screens are very noisy and calculation of FDR led to almost zero signal. This is a prevalent opinion at the RNAi screen field and it was once again confirmed in this work.

- Threshold selection: As explained in section 5.5, RankSLP had experimented with cutting the ranked lists in different thresholds, like top 10%, 20% or 30%. An extensive literature search was made and that is when RNAiCut was detected and applied: the genes in the ranked list are connected with their underlying PPI network, in order to decide the cutting point. More details are provided in sections 3.3 and 4.7. However, the selection of a threshold based on a ranked list per-se is not a trivial task. Therefore, widely used statistical criteria were finally applied (p-value < 0.05 and z-score < -1). As explained in section 5.5, these criteria are not restirctive, since the final result is a combination of numerous methods.

- Multiple shRNAs-to-gene value assignment: For the cases where multiple shRNAs per gene are available, before concluding to the consideration of the best ranking shRNA, the median and the mean were also tested.

- STITCH thresholds. Initially, no confidence threshold at all was applied to the drug-target pairs in STITCH. This doubled the computation time with no gain in accuracy.

- CanSAR dataset. CanSAR is a cancer research and drug discovery knowledgebase, developed by the Computational Biology and Chemogenomics Team, Cancer Research UK Cancer Therapeutics Unit at the ICR[1] [121]. This database was interrogated for drug response IC50 values at its very early version (spring 2012). At this point, the dataset consisted of 67,531 drug profiles on 1,056 cell lines. Analysis of such a huge dataset, for only the tuning of one of the parameters, was taking approximately 5 days (∼100 CPU hours) on one Core of an Intel E7440 CPU, 2.4GHz, 128 GB of memory (available machine at that time). Even with parallelization (R *snow* package), the total time was still around 60 CPU hours. On top of that, at that early version of CanSAR, the drugs were not well annotated and the mutations of the tested cell lines were not provided. A manual search on COSMIC would have taken very long. Overall, processing of the CanSAR dataset whould have caused long delays. Thus, for this version of RankSLP it was ignored, given that the remaining four drug datasets would be able to rescue the most important candidates.

- Aliases-to-SMILES codes mapping: As explained in subsection 4.11.4, a search for global drug identifiers was performed for each of the compound inhibition datasets. A limited amount of SMILES codes was detected for the NCI60 dataset

---

[1] `https://cansar.icr.ac.uk/`

and conversion to InchiKeys was performed. However, this mapping was not used for the final calculation, as explained in 4.11.4.

- Rank Aggregation: Several parameters were tuned for the *RankAggreg()* function application, which is described in section 3.4. Instead of incorporating the rank of best ranking shRNA for each gene, the median, and mean of the top 2 was also considered. Moreover, use of the GA heuristic was attempted, with tuning of many parameters, leading to no conversion. These parameters, were also the ones tested for the CE heuristic: Kendal or Spearman distance, top 5% or top 10% of the ranked lists, inclusion or not of importance weights for each list. Taking into account that RankAggregation takes ∼6 CPU hours, this testing took ∼400 CPU hours.

6.3 Contributions

The current thesis answers the main question: “Which is a set of robust KRAS SLPs?”. This question is biological and this work exhibited how existing computational methods can be used in an innovative manner in order to answer it. The contributions of the current analysis are:

1. The determination of five **new KRAS SLPs** that have not been discovered in the past: BRCA2, COPB2, ERN1, FOS and SPRY1. This brings KRAS-dependent cancers closer to treatment.
2. The **confirmation** of the already known KRAS SLPs: NFKB1 and SMAD1.
3. For the first time, RNAi screens are analyzed integratively, towards SLP finding. The developed pipeline is general and can be applied to synthetic lethal partner detection of other genes as well. Some examples are MYC, p53, Retinoblastoma protein (Rb), protein kinases (e.g. mTOR) and at least 20 more cases [7, 19].
4. It **integrates** existing **datasets** towards the detection of KRAS SLPs, in contradiction with previous studies which involved generation of new data.
    - To our knowledge, this is the first **purely computational** approach towards SLP detection. It is also an implication that in some biology problems, there may be no need for generation of new data. This is much more ethical (less animals killed), and has the advantage of decreased time and costs.
    - Integration of RNAi screens and of external data leads to results with broader applicability.

5. The current approach differs from state-of-the-art methods in the confirmation of its candidates by **external data**. These data form a confident test set, coming especially from high-throughput compound inhibition screens and network information. This is a novel way of approaching the most critical inherent RNAi screen problem, which is noise. The real signal of an RNAi screen is so low that cross-validation is not reliable. External data ensure robustness. In this setting, external data detected NFKB1, which is already a GSG. They also confirmed BRCA2, ERN1, FOS, HDAC9, NCOR1, PAX6, TOP1 (section 4.11) and CTNNA1, ERN1, GSG2, TTK (section 4.12), which are all associated with at least one hallmark of cancer (Figure 5.1).

6. This thesis exhibited how different rankings and their local optima relate. Meta-analysis of rankings leads to global optima solutions.

7. This is the first comparison between RIGER and RSA on real datasets. As shown in subsection 5.3.2, RIGER performs better than RSA in terms of retrieval of Gold Standard Genes. At its top level (10-20%) of recall, it has 60% larger precision. This is in accordance with the amount of publications that reference it: RIGER has 158 references on Google Scholar, while RSA has 70 (as of March 2014).

8. Development of a new method which examines the effects of the inclusion of a random or extra screen (sections 3.5, 4.14).

9. Until now, RNAi screening analysis was based only on the best ranking shRNA or the mean of the first two shRNAs. However, due to noise in RNAi screening, pre-selection of data may lead to loss in accuracy. Thus, RankSLP implements an additional approach that does not have to pre-decide on which shRNAs to include in further ranking; it examines all of them. This method selects its final hits based on the frequency of the hit shRNAs and on statistical criteria (section 4.6). It then integrates these findings with the ones from the rest methods. As a result, it detects two GSGs, SMAD1 and NFKB1, and overall six out of the seven most potential candidates, in which RankSLP concludes.

10. Enriched/top genes are connected in a PPI network, as shown in Chapter Results.

11. After significant tuning, it was *shown* that the 10% percentile is an *acceptable* threshold. However, this was not proved.

The concentrated, final code of RankSLP is ∼2,200 lines, but if all testing described in section 6.2 is considered, this number is doubled. The total calculation time is ∼18 CPU hours on one Core of an Intel X5650 CPU, 2.83Hz, 96 GB of memory.

Suggestions:

1. For proper statistical analysis, more replicates are needed. Thus, biologists are advised to perform multiple experiments, when possible, and computer scientists to opt for datasets with a sufficient sample size.
2. A guideline for biologists, resulting from the current work, could be the performance of synthetic lethal experiments on isogenic cell lines, or, at least, on cell lines with the minimum possible variation. This will ensure "clearer" phenotype and gene signatures.
3. At this point, the obvious should be stated: Closer collaboration of biologists with computer scientists is important. If the data are in a proper format they can quickly and robustly be analyzed. It is not trivial for a bioinformatician to try to "transform" the already generated data and to understand all the details of an experiment, without the guidance of its performer.

6.4 Possible extensions

This work covers a lot of aspects of integrative analysis of high throughput datasets, towards the detection of potent KRAS SLPs. However, there are some parts that could be further investigated in the future and could make this study more complete. These parts are discussed in what follows.

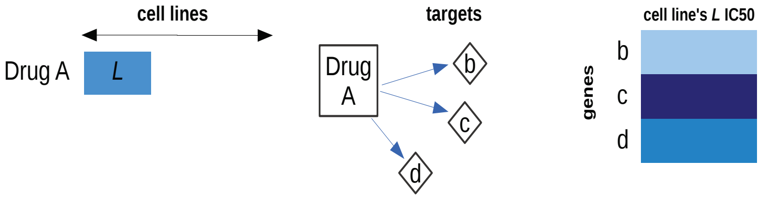


Figure 6.1: This figure is a toy example for the exhibition of a drug's action. A drug D can be effective against a cell line L, translated in a low IC50 for the cell line. Drug D has 3 target genes b,c and d. The sensitivity that L exposes against drug D, is a combinatorial effect of D on its targets. The extend to which each gene's profile contributes to the low *IC50* of L is depicted by different coloring, with darker levels corresponding to stronger effect. The overall effect on L is a mixture of the three shades of blue. There are many ways that the genes' profiles can influence the final drug response phenotype.

- In the comparison with external chemical genomics screens, it is assumed that there are no relationships among the targets of a drug. However, the real picture is that a drug has multiple targets which can interact with each other. This is schematically shown in figure 6.1.

For comparing a viability/apoptosis score from an RNAi screen with an IC50 value from a drug screen, two main approaches that can be followed are suggested:
1. Calculation of the average of the multiple gene entries in the RNAi screen and comparison of this value with the IC50 value of the drug screen. This, however, assumes that the effect of the drug in the drug screen is averaged across its targets, which is not necessarily the case. A drug can have different efficiency on each of its targets. Moreover, the targets of the same drug may interact with each other when targeted. In addition to that, the shRNAs against the same gene in an RNAi screen, are not similarly effective on it. For example, one shRNA may have 70% knock-down efficiency and another 90%.
2. Consideration of a drug's IC50 value from one drug screen, as many times as the number of its targets. Each IC50 will be regressed against one z-score from the RNAi screen, each one corresponding to each of the targets. However, this violates the basic regression assumption which is the existence of independent measurements.
These are points that could be further investigated in the future.

- Another extension in the topic of drug screen analysis is that the updated CanSAR dataset could be included in the pipeline.

- A suggestion is that better annotation of drug datasets, with inclusion of global and unique identifiers (e.g. SMILES codes), should be performed.

- Regarding additional ranking methods that can be applied, the SSMD method presented in section 5.3 seems very promising, provided that special information and additional data on how each screening is performed are accessible.

- In the topic of rankings and more specifically rank aggregation of top n elements of k lists, the followed procedure *RankAggreg* from R with the choice of CE as heuristic and Spearman as the metric for list distance calculation, is $O(n^k) + O(n) = O(n^k)$. This is already a significant decrease from the original complexity time that Kemeny optimization requires (NP-hard). Since in the current case this method is applied on only three lists and still takes around five hours to be computed for approximately the top 100 genes of the lists, there is much room for improvement with the development of quicker algorithms and algorithms that have robust results when they aggregate a larger top-subset of the input lists.

- The synteny of the detected genes could be investigated and evidence from other species could be incorporated. Maybe synteny analysis will reveal some candidates in the proximity of the genes selected by RankSLP.

Last but not least, since this analysis was able to detect a set of well-performing candidate KRAS SLPs on three datasets, proper experimental validation is encouraged. Given the results of this thesis, experimental validation seems feasible and may boost the quality of this study. If successful, it may lead to the development of new therapeutic targets of KRAS mutant dependent cancers.

Appendix A

Appendix

A.1 siRNAs and shRNAs

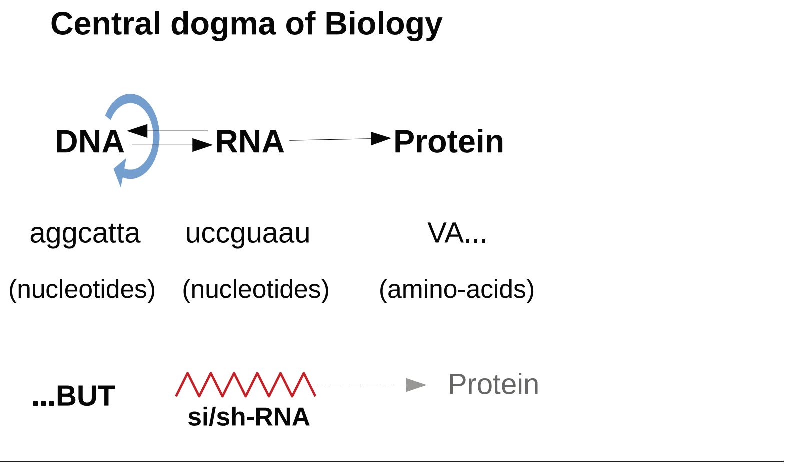


Figure A.1: The central dogma of Biology: DNA can replicate and can also transcribe to RNA. RNA can reverse transcribe and translate to protein. siRNAs or shRNAs are small RNA sequences that are specific to a gene of interest and can turn off its translation. Please note that the amino acids shown on the schematic are just example amino acids and they don't correspond to the preceding nucleotide sequences.

A short hairpin RNA (shRNA) is an RNA sequence, having a characteristic hairpin turn, which can be used to silence target gene expression via RNA interference (RNAi). shRNA's length varies from 19 to 29 base pairs. The shRNA of interest is transcribed by a plasmid vector and it can also be packaged in viruses. The plasmid is used in transfection of human cell lines and the viruses are used for infection of the cell lines. It is then transcribed to a one-stranded small RNA that is loaded into the RNA-induced silencing complex (RISC) where it is pre-processed and "unwanted" parts are cleaved. The remaining part can then bind to target mRNAs and, in turn, silence their translation to proteins.

Small interfering RNA (siRNA) on the other hand is an already processed, 20-25 base pairs long double stranded RNA, stemming from shRNA or dsRNA. It is named after its action, which is to *interfere* with RNA that is transcribed from specific DNA parts of interest (genes) and silence their expression. As a result, the intended protein is not translated. The procedure by which an shRNA or dsRNA is transformed to an siRNA involves the action of the enzyme Dicer, also known as 'endoribonuclease Dicer' or 'helicase with RNase motif'. Dicer cleaves the shRNAs or dsRNAs to double stranded siRNAs. It also helps towards the activation of the RNA-induced silencing complex (RISC), which is essential for RNA interference realization. Its function is exposed in figure A.2. The activity of siRNAs depends on its binding activity to RISC. Their function is similar to microRNAs; their difference is that microRNAs have a complementary, while siRNAs have specific sequence to the DNA of interest.

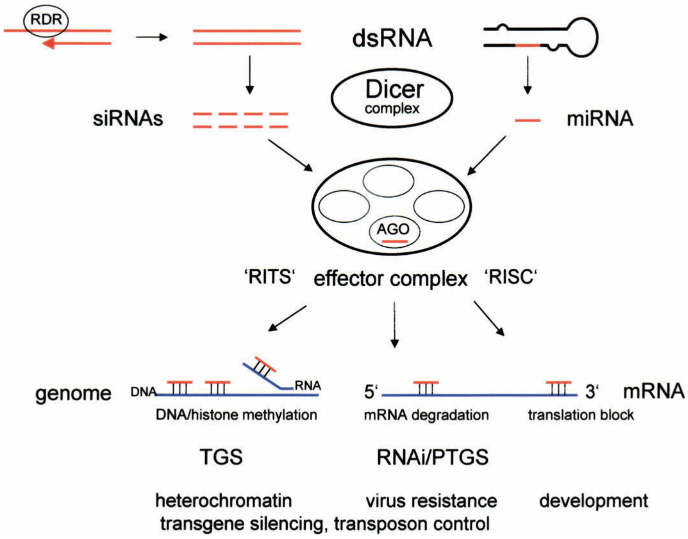


FIGURE A.2: Short RNAs derived from Dicer cleavage of dsRNA are incorporated into multiprotein effector complexes, such as RISC and RITS (RNA-induced initiation of TGS) to target mRNA degradation (RNAi/PTGS), translation inhibition, or TGS and genome modifications. Figure and caption are taken from [122]

The use of both siRNAs and shRNAs for gene silencing seems to be a great tool towards personalized cancer therapy. Despite its promising nature, it is very challenging, and should be performed with caution. The reason for that is that they both exhibit different effectiveness on different types of cells. They also both have Off Target Effects (OTEs), meaning unintended silencing of genes with close sequence similarity to the gene that actually needs to be blocked. Many reviews exist on their function and whether the one or the other is better to use [123], [124]. Despite profound research, they are inconclusive, since both siRNAs and shRNAs have advantages and disadvantages. For example, siRNA transfection is said to be more effective than shRNA, but, on the other hand, more OTEs are observed with siRNAs. A common observation is that when the quantity of both the DNA and the reagent is increased, they both lead to a clearer phenotype. Overall, it seems that there is no 'best' solution, and but always depends on the types of cells, of the targeted genes and of the overall experimental design.

Bibliography

[1] R. Hesketh. *Introduction to Cancer Biology.* Cambridge University Press, 2012.

[2] Ana M Mendes-Pereira, David Sims, Tim Dexter, Kerry Fenwick, Ioannis Assiotis, Iwanka Kozarewa, Costas Mitsopoulos, Jarle Hakas, Marketa Zvelebil, Christopher J Lord, and Alan Ashworth. Genome-wide functional screen identifies a compendium of genes affecting sensitivity to tamoxifen. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8):2730–5, 2012.

[3] K. R. Loeb. Significance of multiple mutations in cancer. *Carcinogenesis*, 21(3): 379–385, 2000.

[4] Bert Vogelstein and Kenneth W Kinzler. Cancer genes and the pathways they control. *Nature medicine*, 10(8):789–99, 2004.

[5] Paul M Campbell and Channing J Der. Oncogenic Ras and its role in tumor cell invasion and metastasis. *Seminars in cancer biology*, 14(2):105–14, 2004.

[6] Anurag Singh, Patricia Greninger, Daniel Rhodes, Louise Koopman, Sheila Violette, Nabeel Bardeesy, and Jeff Settleman. A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. *Cancer cell*, 15(6):489–500, 2009.

[7] Ulrich H Weidle, Daniela Maisel, and Dirk Eick. Synthetic lethality-based targets for discovery of new cancer therapeutics. *Cancer genomics & proteomics*, 8(4): 159–71, 2011.

[8] Johannes L Bos. ras Oncogenes in Human Cancer : A Review ras Oncogenes in Human Cancer : A Review1. *Cancer Research*, 49:4682–4689, 1989.

[9] C. B. Bridges and K. S. Brehme. *The mutants of Drosophila melanogaster.* Carnegie Inst Wash Publ, 1944.

[10] T Dobzhansky. Genetics of Natural Populations. Xiii. Recombination and Variability in Populations of Drosophila Pseudoobscura. *Genetics*, 31(3):269–90, 1946.

[11] Sebastian M B Nijman. Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS letters*, 585(1):1–6, 2011.

[12] Elisa Ferrari, Chiara Lucca, and Marco Foiani. A lethal combination for cancer cells: synthetic lethality screenings for drug discovery. *European journal of cancer (Oxford, England : 1990)*, 46(16):2889–95, 2010.

[13] William G. Kaelin. The Concept of Synthetic Lethality in the Context of Anticancer Therapy. *Nature Reviews Cancer*, 5(9):689–698, 2005.

[14] A Persidis. Cancer multidrug resistance. *Nature biotechnology*, 17(1):94–5, 1999.

[15] Frank L Meyskens and Eugene W Gerner. Back to the future: mechanism-based, mutation-specific combination chemoprevention with a synthetic lethality approach. *Cancer prevention research (Philadelphia, Pa.)*, 4(5):628–32, 2011.

[16] K. Garber. Synthetic Lethality: Killing Cancer With Cancer. *CancerSpectrum Knowledge Environment*, 94(22):1666–1668, 2002.

[17] H Christian Reinhardt, Hai Jiang, Michael T Hemann, and Michael B Yaffe. Exploiting synthetic lethal interactions for targeted cancer therapy. *Cell cycle (Georgetown, Tex.)*, 8(19):3112–9, 2009.

[18] Karen A Gelmon, Marc Tischkowitz, Helen Mackay, Kenneth Swenerton, André Robidoux, Katia Tonkin, Hal Hirte, David Huntsman, Mark Clemons, Blake Gilks, Rinat Yerushalmi, Euan Macpherson, James Carmichael, and Amit Oza. Olaparib in patients with recurrent high-grade serous or poorly differentiated ovarian carcinoma or triple-negative breast cancer: a phase 2, multicentre, open-label, non-randomised study. *The lancet oncology*, 12(9):852–61, 2011.

[19] John M Furgason and El Mustapha Bahassi. Targeting DNA repair mechanisms in cancer. *Pharmacology & therapeutics*, 137(3):298–308, 2013.

[20] Amy Young, Jesse Lyons, Abigail L Miller, Vernon T Phan, Irma Rangel Alarcón, and Frank McCormick. Ras signaling and therapies. *Advances in cancer research*, 102:1–17, 2009.

[21] Julian Downward. Targeting RAS signalling pathways in cancer therapy. *Nature reviews. Cancer*, 3(1):11–22, 2003.

[22] Jacob John, Roland Sohmen, Juergen Feuerstein, Rosita Linke, Alfred Wittinghofer, and Roger S. Goody. Kinetics of interaction of nucleotides with nucleotide-free H-ras p21. *Biochemistry*, 29(25):6058–6065, 1990.

[23] Jonathan M Ostrem, Ulf Peters, Martin L Sos, James A Wells, and Kevan M Shokat. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature*, 503(7477):548–51, 2013.

[24] David a Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Fröhling, Edmond M Chan, Martin L Sos, Kathrin Michel, Craig Mermel, Serena J Silver, Barbara a Weir, Jan H Reiling, Qing Sheng, Piyush B Gupta, Raymond C Wadlow, Hanh Le, Sebastian Hoersch, Ben S Wittner, Sridhar Ramaswamy, David M Livingston, David M Sabatini, Matthew Meyerson, Roman K Thomas, Eric S Lander, Jill P Mesirov, David E Root, D Gary Gilliland, Tyler Jacks, and William C Hahn. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–12, 2009.

[25] Ji Luo, Michael J Emanuele, Danan Li, Chad J Creighton, Michael R Schlabach, Thomas F Westbrook, Kwok-kin Wong, and Stephen J Elledge. A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. 137(5):835–848, 2009.

[26] Michael Steckel, Miriam Molina-Arcas, Britta Weigelt, Michaela Marani, Patricia H Warne, Hanna Kuznetsov, Gavin Kelly, Becky Saunders, Michael Howell, Julian Downward, and David C Hancock. Determination of synthetic lethal interactions in KRAS oncogene-dependent cancer cells reveals novel therapeutic targeting strategies. *Cell research*, 22(8):1227–45, 2012.

[27] Y Wang, V N Ngo, M Marani, Y Yang, G Wright, L M Staudt, and J Downward. Critical role for transcriptional repressor Snail2 in transformation by oncogenic RAS in colorectal carcinoma cells. *Oncogene*, 29(33):4658–70, 2010.

[28] Claudia Scholl, Stefan Fröhling, Ian F Dunn, Anna C Schinzel, David a Barbie, So Young Kim, Serena J Silver, Pablo Tamayo, Raymond C Wadlow, Sridhar Ramaswamy, Konstanze Döhner, Lars Bullinger, Peter Sandy, Jesse S Boehm, David E Root, Tyler Jacks, William C Hahn, and D Gary Gilliland. Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell*, 137(5):821–34, 2009.

[29] Richard Robinson. RNAi therapeutics: how likely, how soon? *PLoS biology*, 2: E28, 2004.

[30] C J Lord, S a Martin, and A Ashworth. RNA interference screening demystified. *Journal of clinical pathology*, 62(3):195–200, 2009.

[31] Christophe J Echeverri and Norbert Perrimon. High-throughput RNAi screening in cultured cells: a user's guide. *Nature reviews. Genetics*, 7(5):373–84, 2006.

[32] J Mullenders and R Bernards. Loss-of-function genetic screens as a tool to improve the diagnosis and treatment of cancer. *Oncogene*, 28(50):4409–20, 2009.

[33] L. H. Hartwell. Integrating Genetic Approaches into the Discovery of Anticancer Drugs. *Science*, 278(5340):1064–1068, 1997.

[34] Natasha J Caplen. RNAi as a gene therapy approach. *Expert opinion on biological therapy*, 3(4):575–86, 2003.

[35] Olivia Perwitasari, Abhijeet Bakre, S Mark Tompkins, and Ralph a Tripp. siRNA Genome Screening Approaches to Therapeutic Drug Repositioning. *Pharmaceuticals (Basel, Switzerland)*, 6(2):124–60, 2013.

[36] Bhavneet Bhinder and Hakim Djaballah. A decade of RNAi screening: Too much hay and very few needles. *Drug Discovery World*, 14(3), 2013. ISSN 14694344.

[37] Shuo Gu, Lan Jin, Yue Zhang, Yong Huang, Feijie Zhang, Paul N Valdmanis, and Mark A Kay. The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo. *Cell*, 151(4):900–11, 2012.

[38] Jong-Eun Park, Inha Heo, Yuan Tian, Dhirendra K Simanshu, Hyeshik Chang, David Jee, Dinshaw J Patel, and V Narry Kim. Dicer recognizes the 5' end of RNA for efficient and accurate processing. *Nature*, 475(7355):201–5, 2011.

[39] Aimee L Jackson, Steven R Bartz, Janell Schelter, Sumire V Kobayashi, Julja Burchard, Mao Mao, Bin Li, Guy Cavet, and Peter S Linsley. Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, 21(6):635–638, 2003.

[40] Andrea Franceschini, Roger Meier, Alain Casanova, Saskia Kreibich, Neha Daga, Daniel Andritschke, Sabrina Dilling, Pauli Rämö, Mario Emmenlauer, Andreas Kaufmann, Raquel Conde-Álvarez, Shyan Huey Low, Lucas Pelkmans, Ari Helenius, Wolf-Dietrich Hardt, Christoph Dehio, and Christian von Mering. Specific inhibition of diverse pathogens in human cells by synthetic microRNA-like oligonucleotides inferred from RNAi screens. *Proceedings of the National Academy of Sciences of the United States of America*, 111(12):4548–53, 2014.

[41] Tuoping Luo, Kristina Masson, Jacob D Jaffe, Whitney Silkworth, Nathan T Ross, Christina A Scherer, Claudia Scholl, Stefan Fröhling, Steven A Carr, Andrew M Stern, Stuart L Schreiber, and Todd R Golub. STK33 kinase inhibitor BRD-8899

has no effect on KRAS-dependent cancer cell viability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8):2860–5, 2012.

[42] Carol Babij, Yihong Zhang, Robert J Kurzeja, Anke Munzli, Amro Shehabeldin, Manory Fernando, Kim Quon, Paul D Kassner, Astrid A Ruefli-Brasse, Vivienne J Watson, Flordeliza Fajardo, Angela Jackson, James Zondlo, Yu Sun, Aaron R Ellison, Cherylene A Plewa, Miguel Tisha San, John Robinson, John McCarter, Ralf Schwandner, Ted Judd, Josette Carnahan, and Isabelle Dussault. STK33 kinase activity is nonessential in KRAS-dependent cancer cells. *Cancer research*, 71(17):5818–26, 2011.

[43] B Vangamudi, A. E. Ayres, J. P. Burke, A. G. Waterson, O. W. Rossanese1, and Fesik S. W. Evaluation of TBK1 as a novel cancer target in the K-Ras pathway . *Cancer Research*, 72(8), 2012.

[44] S A Watt, C Pourreyron, K Purdie, C Hogan, C L Cole, N Foster, N Pratt, J-C Bourdon, V Appleyard, K Murray, A M Thompson, X Mao, C Mein, L Bruckner-Tuderman, A Evans, J A McGrath, C M Proby, J Foerster, I M Leigh, and A P South. Integrative mRNA profiling comparing cultured primary cells with clinical samples reveals PLK1 and C20orf20 as therapeutic targets in cutaneous squamous cell carcinoma. *Oncogene*, 30(46):4666–77, 2011.

[45] Yan Ding, Dan Huang, Zhongfa Zhang, Josh Smith, David Petillo, Brendan D Looyenga, Kristin Feenstra, Jeffrey P Mackeigan, Kyle A Furge, and Bin T Teh. Combined gene expression profiling and RNAi screening in clear cell renal cell carcinoma identify PLK1 and other therapeutic kinase targets. *Cancer research*, 71(15):5225–34, 2011.

[46] Aparna V Sarthy, Susan E Morgan-Lappe, Dorothy Zakula, Lawrence Vernetti, Mark Schurdak, Jeremy C L Packer, Mark G Anderson, Senji Shirasawa, Takehiko Sasazuki, and Stephen W Fesik. Survivin depletion preferentially reduces the survival of activated K-Ras-transformed cells. *Molecular cancer therapeutics*, 6(1): 269–76, 2007.

[47] Kristina A Cole, Jonathan Huggins, Michael Laquaglia, Chase E Hulderman, Mike R Russell, Kristopher Bosse, Sharon J Diskin, Edward F Attiyeh, Rachel Sennett, Geoffrey Norris, Marci Laudenslager, Andrew C Wood, Patrick A Mayes, Jayanti Jagannathan, Cynthia Winter, Yael P Mosse, and John M Maris. RNAi screen of the protein kinome identifies checkpoint kinase 1 (CHK1) as a therapeutic target in neuroblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 108(8):3336–41, 2011.

[48] Min Zheng, Susan E Morgan-Lappe, Jie Yang, Katrina M Bockbrader, Deepika Pamarthy, Dafydd Thomas, Stephen W Fesik, and Yi Sun. Growth inhibition and radiosensitization of glioblastoma and lung cancer cells by small interfering RNA silencing of tumor necrosis factor receptor-associated factor 2. *Cancer research*, 68 (18):7570–8, 2008.

[49] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F Berger, John E Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H Engels, Jill Cheng, Guoying K Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P Mesirov, Stacey B Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E Myer, Barbara L Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L Harris, Matthew Meyerson, Todd R Golub, Michael P Morrissey, William R Sellers, Robert Schlegel, and Levi A Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–7, 2012.

[50] Silvia Domcke, Rileen Sinha, Douglas a Levine, Chris Sander, and Nikolaus Schultz. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature communications*, 4:2126, 2013.

[51] Cynthia Dwork. Rank Aggregation Revisited. 2001.

[52] Maria Arcila, Christopher Lau, Khedoudja Nafa, and Marc Ladanyi. Detection of KRAS and BRAF mutations in colorectal carcinoma roles for high-sensitivity locked nucleic acid-PCR sequencing and broad-spectrum mass spectrometry genotyping. *The Journal of molecular diagnostics : JMD*, 13(1):64–73, 2011.

[53] Michael R Schlabach, Ji Luo, Nicole L Solimini, Guang Hu, Qikai Xu, Mamie Z Li, Zhenming Zhao, Agata Smogorzewska, Mathew E Sowa, Xiaolu L Ang, Thomas F Westbrook, Anthony C Liang, Kenneth Chang, Jennifer a Hackett, J Wade Harper, Gregory J Hannon, and Stephen J Elledge. Cancer proliferation gene discovery through functional genomics. *Science (New York, N.Y.)*, 319:620–4, 2008.

[54] Biao Luo, Hiu Wing Cheung, Aravind Subramanian, Tanaz Sharifnia, Michael Okamoto, Xiaoping Yang, Greg Hinkle, Jesse S Boehm, Rameen Beroukhim, Barbara a Weir, Craig Mermel, David a Barbie, Tarif Awad, Xiaochuan Zhou, Tuyen

Nguyen, Bruno Piqani, Cheng Li, Todd R Golub, Matthew Meyerson, Nir Hacohen, William C Hahn, Eric S Lander, David M Sabatini, and David E Root. Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20380–5, 2008.

[55] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, and Benjamin L Ebert. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 2005.

[56] Renate Koenig, Chih-yuan Chiang, Buu P Tu, Renate Ko, S Frank Yan, Paul D Dejesus, Angelica Romero, Tobias Bergauer, Anthony Orth, Ute Krueger, Yingyao Zhou, and Sumit K Chanda. A probability-based approach for the analysis of large-scale RNAi screens. *Nature ...*, 4(10):847–849, 2007.

[57] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. 2006.

[58] Irene M Kaplow, Rohit Singh, Adam Friedman, Chris Bakal, Norbert Perrimon, and Bonnie Berger. RNAiCut: automated detection of significant genes from functional genomic screens. *Nature methods*, 6(7):476–7, 2009.

[59] Vasyl Pihur, Susmita Datta, and Somnath Datta. RankAggreg, an R package for weighted rank aggregation. *BMC bioinformatics*, 10(1):62, 2009.

[60] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Artificial Intelligence. Addison-Wesley, 1989. ISBN 9780201157673. URL http://books.google.de/books?id=3_RQAAAAMAAJ.

[61] Vasyl Pihur, Susmita Datta, and Somnath Datta. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics (Oxford, England)*, 23(13):1607–15, 2007.

[62] Marta Puyol, Alberto Martín, Pierre Dubus, Francisca Mulero, Pilar Pizcueta, Gulfaraz Khan, Carmen Guerra, David Santamaría, and Mariano Barbacid. A synthetic lethal interaction between K-Ras oncogenes and Cdk4 unveils a therapeutic strategy for non-small cell lung carcinoma. *Cancer cell*, 18(1):63–73, 2010.

[63] Ryan B Corcoran, Katherine a Cheng, Aaron N Hata, Anthony C Faber, Hiromichi Ebi, Erin M Coffee, Patricia Greninger, Ronald D Brown, Jason T Godfrey, Travis J Cohoon, Youngchul Song, Eugene Lifshits, Kenneth E Hung, Toshi Shioda, Dora Dias-Santagata, Anurag Singh, Jeffrey Settleman, Cyril H Benes,

Mari Mino-Kenudson, Kwok-Kin Wong, and Jeffrey a Engelman. Synthetic lethal interaction of combined BCL-XL and MEK inhibition promotes tumor regressions in KRAS mutant cancer models. *Cancer cell*, 23(1):121–8, 2013.

[64] Nicole Bennardo, Anita Cheng, Nick Huang, and Jeremy M Stark. Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLoS genetics*, 4(6), 2008.

[65] Oren Gilad, Barzin Y Nabet, Ryan L Ragland, David W Schoppy, Kevin D Smith, Amy C Durham, and Eric J Brown. Combining ATR suppression with oncogenic Ras synergistically increases genomic instability, causing synthetic lethality or tumorigenesis in a dosage-dependent manner. *Cancer research*, 70(23):9693–702, 2010.

[66] Matthew J Sale and Simon J Cook. The BH3 mimetic ABT-263 synergizes with the MEK1/2 inhibitor selumetinib/AZD6244 to promote BIM-dependent tumour cell death and inhibit acquired resistance. *The Biochemical journal*, 450(2):285–94, 2013.

[67] J Ho and a Bretscher. Ras regulates the polarity of the yeast actin cytoskeleton through the stress response pathway. *Molecular biology of the cell*, 12(6):1541–55, 2001.

[68] H. Hattori, F. Skoulidis, P. Russell, and a. R. Venkitaraman. Context Dependence of Checkpoint Kinase 1 as a Therapeutic Target for Pancreatic Cancers Deficient in the BRCA2 Tumor Suppressor. *Molecular Cancer Therapeutics*, 10(4):670–678, 2011.

[69] Zubaidah M Ramdzan, Charles Vadnais, Ranjana Pal, Guillaume Vandal, Chantal Cadieux, Lam Leduy, Sayeh Davoudi, Laura Hulea, Lu Yao, Anthony N Karnezis, Marilène Paquet, David Dankort, and Alain Nepveu. RAS Transformation Requires CUX1-Dependent Repair of Oxidative DNA Damage. *PLoS biology*, 12(3), 2014.

[70] Xiangwei Wu and Scott M Lippman. An intermittent approach for cancer chemoprevention. *Nature reviews. Cancer*, 11(12):879–85, 2011.

[71] Madhu S Kumar, David C Hancock, Miriam Molina-Arcas, Michael Steckel, Phillip East, Markus Diefenbacher, Elena Armenteros-Monterroso, François Lassailly, Nik Matthews, Emma Nye, Gordon Stamp, Axel Behrens, and Julian Downward. The GATA2 transcriptional network is requisite for RAS oncogene-driven non-small cell lung cancer. *Cell*, 149(3):642–55, 2012.

[72] Song Shen, Chong-Qiong Mao, Xian-Zhu Yang, Xiao-Jiao Du, Yang Liu, Yan-Hua Zhu, and Jun Wang. Cationic Lipid-Assisted Polymeric Nanoparticle Mediated GATA2 siRNA Delivery for Synthetic Lethal Therapy of KRAS Mutant Non-Small-Cell Lung Carcinoma. *Molecular pharmaceutics*, 2014.

[73] Niki Karachaliou, Clara Mayo, Carlota Costa, Ignacio Magrí, Ana Gimenez-Capitan, Miguel Angel Molina-Vila, and Rafael Rosell. KRAS mutations in lung cancer. *Clinical lung cancer*, 14(3):205–14, 2013.

[74] Lisa M Nilsson, Tacha Zi Plym Forshell, Sara Rimpi, Christiane Kreutzer, Walter Pretsch, Georg W Bornkamm, and Jonas a Nilsson. Mouse genetics suggests cell-context dependency for Myc-regulated metabolic enzymes during tumorigenesis. *PLoS genetics*, 8(3), 2012.

[75] Anurag Singh, Michael F Sweeney, Min Yu, Alexa Burger, Patricia Greninger, Cyril Benes, Daniel a Haber, and Jeff Settleman. TAK1 inhibition promotes apoptosis in KRAS-dependent colon cancers. *Cell*, 148(4):639–50, 2012.

[76] Asami Takashima and Douglas V Faller. Targeting the RAS oncogene. *Expert opinion on therapeutic targets*, 17(5):507–31, 2013.

[77] Alice T Shaw, Monte M Winslow, Margaret Magendantz, Chensi Ouyang, James Dowdle, Aravind Subramanian, Timothy A Lewis, Rebecca L Maglathin, Nicola Tolliday, and Tyler Jacks. Selective killing of K-ras mutant cancer cells by small molecule inducers of oxidative stress. *Proceedings of the National Academy of Sciences of the United States of America*, 108(21):8773–8, 2011.

[78] Zhihong Chen, Lora W Forman, Robert M Williams, and Douglas V Faller. Protein kinase C-delta inactivation inhibits the proliferation and survival of cancer stem cells in culture and in vivo. *BMC cancer*, 14:90, 2014.

[79] Phuoc T Tran, Emelyn H Shroff, Timothy F Burns, Saravanan Thiyagarajan, Sandhya T Das, Tahera Zabuawala, Joy Chen, Yoon-Jae Cho, Richard Luong, Pablo Tamayo, Tarek Salih, Khaled Aziz, Stacey J Adam, Silvestre Vicent, Carsten H Nielsen, Nadia Withofs, Alejandro Sweet-Cordero, Sanjiv S Gambhir, Charles M Rudin, and Dean W Felsher. Twist1 suppresses senescence programs and thereby accelerates and maintains mutant Kras-induced lung tumorigenesis. *PLoS genetics*, 8(5):e1002650, 2012.

[80] Wan Seok Yang and Brent R Stockwell. Synthetic lethal screening identifies compounds activating iron-dependent, nonapoptotic cell death in oncogenic-RAS-harboring cancer cells. *Chemistry & biology*, 15(3):234–45, 2008.

[81] Silvia Licciulli and JL Kissil. WT1: a weak spot in KRAS-induced transformation. *The Journal of clinical investigation*, 120(11):9–12, 2010.

[82] N. E. Faulkner, M. M. Da Silva, R. A. Heim, B. C. Horten, E. M. Rohlfs, L. S. Rosenblum, B. A. Allitto, and D. A. Sirko-Osadsa. KRAS mutation analyses of more than 16,500 colorectal carcinomas. Meeting: 2010 Molecular Markers, General Poster Session, Abstract Number:96, 2010.

[83] Bingliang Fang. Development of Synthetic Lethality Anticancer Therapeutics. *Journal of medicinal chemistry*, 2014.

[84] Thomas Kelder, Martijn P van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, 2012. doi: 10.1093/nar/gkr1074.

[85] D Lancet, G Stelzer, Y Golan, and A Rinon. Gene Trends: On Muscle, Fat, and Brain. *Genetic Engineering & Biotechnology News*, 2013.

[86] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, Sridhar Ramaswamy, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Cyril Benes, Ultan McDermott, and Mathew J Garnett. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(Database issue):D955–61, 2013.

[87] Robert H Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature reviews. Cancer*, 6(10):813–23, 2006.

[88] William C Reinhold, Margot Sunshine, Hongfang Liu, Sudhir Varma, Kurt W Kohn, Joel Morris, James Doroshow, and Yves Pommier. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer research*, 72(14):3499–511, 2012.

[89] Uma T Shankavaram, Sudhir Varma, David Kane, Margot Sunshine, Krishna K Chary, William C Reinhold, Yves Pommier, and John N Weinstein. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC genomics*, 10(1):277, 2009.

[90] S Bamford, E Dawson, S Forbes, J Clements, R Pettett, a Dogan, a Flanagan, J Teague, P a Futreal, M R Stratton, and R Wooster. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer*, 91(2):355–8, 2004.

[91] Michael Kuhn, Damian Szklarczyk, Andrea Franceschini, Christian von Mering, Lars Juhl Jensen, and Peer Bork. STITCH 3: zooming in on protein-chemical interactions. *Nucleic acids research*, 40(Database issue):D876–80, 2012.

[92] Mark E. Burkard. Integrating the NCI-60 Data with “Omics” for Drug Discovery. *Drug Development Research*, 73(7):420–429, 2012. doi: 10.1002/ddr.21033.

[93] Bing Zhang, Stefan Kirov, and Jay Snoddy. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*, 33(Web Server issue):W741–8, 2005.

[94] Jing Wang, Dexter Duncan, Zhiao Shi, and Bing Zhang. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research*, 41(Web Server issue):W77–83, 2013.

[95] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Ozgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, 39(Database issue):D685–90, 2011.

[96] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, 2000.

[97] G Manning, D B Whyte, R Martinez, T Hunter, and S Sudarsanam. The protein kinase complement of the human genome. *Science (New York, N.Y.)*, 298(5600): 1912–34, 2002.

[98] Diane D. Shao, Wen Xue, Elsa B. Krall, Arjun Bhutkar, Federica Piccioni, Xiaoxing Wang, Anna C. Schinzel, Sabina Sood, Joseph Rosenbluh, Jong W. Kim, Yaara Zwang, Thomas M. Roberts, David E. Root, Tyler Jacks, and William C. Hahn. KRAS and YAP1 Converge to Regulate EMT and Tumor Survival. *Cell*, 2014.

[99] Iwona Stelniec-Klotz, Stefan Legewie, Oleg Tchernitsa, Franziska Witzel, Bertram Klinger, Christine Sers, Hanspeter Herzel, Nils Blüthgen, and Reinhold Schäfer. Reverse engineering a hierarchical regulatory network downstream of oncogenic KRAS. *Molecular systems biology*, 8:601, 2012.

[100] Gerben Schaaf, Mohamed Hamdi, Danny Zwijnenburg, Arjan Lakeman, Dirk Geerts, Rogier Versteeg, and Marcel Kool. Silencing of SPRY1 triggers complete

regression of rhabdomyosarcoma tumors carrying a mutated RAS gene. *Cancer research*, 70(2):762–71, 2010.

[101] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011.

[102] Douglas Hanahan, Robert A Weinberg, and San Francisco. The Hallmarks of Cancer Review University of California at San Francisco. 100:57–70, 2000.

[103] Xiaohua Douglas Zhang, Xiting Cindy Yang, Namjin Chung, Adam Gates, Erica Stec, Priya Kunapuli, Dan J Holder, Marc Ferrer, and Amy S Espeseth. Robust statistical methods for hit selection in RNA interference high-throughput screening experiments. *Pharmacogenomics*, 7(3):299–309, 2006.

[104] Namjin Chung, Xiaohua Douglas Zhang, Anthony Kreamer, Louis Locco, Pei-Fen Kuan, Steven Bartz, Peter S Linsley, Marc Ferrer, and Berta Strulovici. Median absolute deviation to improve hit selection for genome-scale RNAi screens. *Journal of biomolecular screening*, 13(2):149–58, 2008.

[105] Amanda Birmingham, Laura M Selfors, Thorsten Forster, David Wrobel, Caleb J Kennedy, Emma Shanks, Javier Santoyo-Lopez, Dara J Dunican, Aideen Long, Dermot Kelleher, Queta Smith, Roderick L Beijersbergen, Peter Ghazal, and Caroline E Shamu. Statistical methods for analysis of high-throughput RNA interference screens. *Nature methods*, 6(8):569–75, 2009.

[106] Juliane Siebourg, Gunter Merdes, Benjamin Misselwitz, Wolf-Dietrich Hardt, and Niko Beerenwinkel. Stability of gene rankings from RNAi screens. *Bioinformatics (Oxford, England)*, 28(12):1612–8, 2012.

[107] Jianping Zhang. *Statistical Modeling for Multiplex RNAi Screen Data Analysis.* PhD thesis, Stony Brook University, 2010.

[108] Asli N Goktug, Su Sien Ong, and Taosheng Chen. GUItars: a GUI tool for analysis of high-throughput RNA interference screening data. *PloS one*, 7(11): e49386, 2012.

[109] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1-3):83–92, 2004.

[110] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–21, 2001.

[111] Barbara Nicke, Julie Bastien, Sophia J Khanna, Patricia H Warne, Victoria Cowling, Simon J Cook, Gordon Peters, Oona Delpuech, Almut Schulze, Katrien Berns, Jasper Mullenders, Roderick L Beijersbergen, René Bernards, Trivadi S Ganesan, Julian Downward, and David C Hancock. Involvement of MINK, a Ste20 family kinase, in Ras oncogene-induced growth arrest in human ovarian surface epithelial cells. *Molecular cell*, 20(5):673–85, 2005.

[112] Twan van Laarhoven, Sander B Nabuurs, and Elena Marchiori. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics (Oxford, England)*, 27(21):3036–43, 2011.

[113] Tapio Pahikkala, Hanna Suominen, and Jorma Boberg. Efficient cross-validation for kernelized least-squares regression with sparse basis expansions. *Machine Learning*, 87(3):381–407, 2012.

[114] Masafumi Toyoshima, Heather L Howie, Maki Imakura, Ryan M Walsh, James E Annis, Aaron N Chang, Jason Frazier, B Nelson Chau, Andrey Loboda, Peter S Linsley, Michele A Cleary, Julie R Park, and Carla Grandori. Functional genomics identifies therapeutic targets for MYC-driven cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24):9545–50, 2012.

[115] Carla Grandori. A high-throughput siRNA screening platform to identify MYC-synthetic lethal genes as candidate therapeutic targets. *Methods in molecular biology (Clifton, N.J.)*, 1012:187–200, 2013.

[116] Yang Gao, Subarma Sinha, and David L. Dill. Identification of Genes that can Selectively Kill Cancer Cells Using Boolean Implications. 2013.

[117] S Eser, a Schnieke, G Schneider, and D Saur. Oncogenic KRAS signalling in pancreatic cancer. *British journal of cancer*, (April):1–6, 2014.

[118] Yuanxiang Wang, Christine E Kaiser, Brendan Frett, and Hong-Yu Li. Targeting mutant KRAS for anticancer therapeutics: a review of novel small molecule modulators. *Journal of medicinal chemistry*, 56(13):5219–30, 2013.

[119] Jeffrey A Engelman, Liang Chen, Xiaohong Tan, Katherine Crosby, Alexander R Guimaraes, Rabi Upadhyay, Michel Maira, Kate McNamara, Samanthi A Perera, Youngchul Song, Lucian R Chirieac, Ramneet Kaur, Angela Lightbown, Jessica Simendinger, Timothy Li, Robert F Padera, Carlos García-Echeverría, Ralph Weissleder, Umar Mahmood, Lewis C Cantley, and Kwok-Kin Wong. Effective use of PI3K and MEK inhibitors to treat mutant Kras G12D and PIK3CA H1047R murine lung cancers. *Nature medicine*, 14(12):1351–6, 2008.

[120] Miriam Molina-Arcas, David C Hancock, Clare Sheridan, Madhu S Kumar, and Julian Downward. Coordinate direct input of both KRAS and IGF1 receptor to activation of PI3 kinase in KRAS-mutant lung cancer. *Cancer discovery*, 3(5): 548–63, 2013.

[121] Krishna C Bulusu, Joseph E Tym, Elizabeth A Coker, Amanda C Schierz, and Bissan Al-Lazikani. canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic acids research*, 42(Database issue):D1040–7, January 2014.

[122] Marjori A Matzke and Antonius J M Matzke. Planting the seeds of a new paradigm. *PLoS biology*, 2(5):E133, 2004.

[123] Dirk Grimm. Small silencing RNAs: state-of-the-art. *Advanced drug delivery reviews*, 61(9):672–703, 2009.

[124] Donald D Rao, John S Vorhies, Neil Senzer, and John Nemunaitis. siRNA vs. shRNA: similarities and differences. *Advanced drug delivery reviews*, 61(9):746–59, 2009.