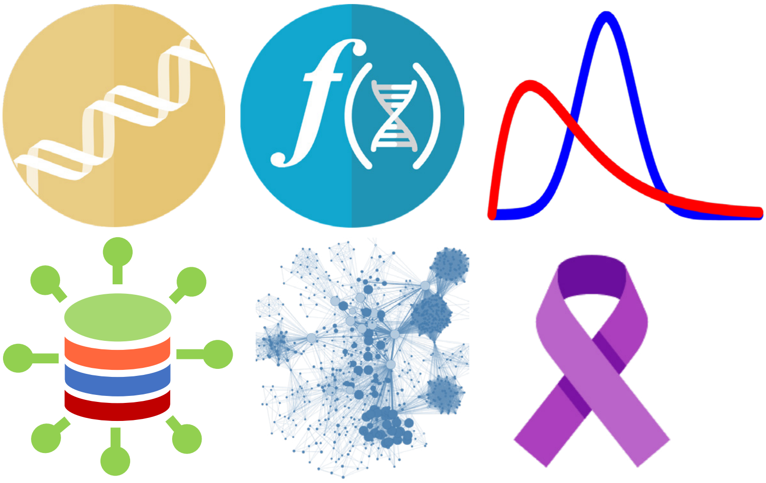
DISSERTATIONES SCHOLAE DOCTORALIS AD SANITATEM INVESTIGANDAM
UNIVERSITATIS HELSINKIENSIS

33/2018

ALOK JAISWAL

Integrative Bioinformatics of Functional and Genomic Profiles for Cancer Systems Medicine



INSTITUTE FOR MOLECULAR MEDICINE FINLAND (FIMM)
FACULTY OF MEDICINE
DOCTORAL PROGRAMME IN INTEGRATIVE LIFE SCIENCE
UNIVERSITY OF HELSINKI

Faculty of Medicine
University of Helsinki
Finland

INTEGRATIVE BIOINFORMATICS OF FUNCTIONAL AND GENOMIC PROFILES FOR CANCER SYSTEMS MEDICINE

Alok Jaiswal

Institute for Molecular Medicine Finland (FIMM)
University of Helsinki
and
Doctoral Program in Integrative Life Science (ILS)
University of Helsinki

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in Lecture Hall 2, Biomedicum Helsinki, Haartmaninkatu 8 on 8th June, 2018 at 12 noon.

Helsinki 2018

Supervisors

Prof. Tero Aittokallio, PhD
Institute for Molecular
Medicine Finland (FIMM),
University of Helsinki,
Helsinki, Finland

Jing Tang, PhD
Institute for Molecular
Medicine Finland (FIMM),
University of Helsinki,
Helsinki, Finland

Thesis Advisory Committee

Brendan Battersby, PhD
Institute of Biotechnology,
University of Helsinki,
Helsinki, Finland

Laura Elo, PhD
Turku Centre for Biotechnology,
University of Turku,
Turku, Finland

Thesis Reviewers

Prof. Garry Wong, PhD
Faculty of Health Science,
University of Macau,
Macau, China

Assoc. prof. Sven Nelander, PhD
Science for Life Laboratory,
Uppsala University,
Uppsala, Sweden

Opponent

Asst. prof. Benjamin Haibe-Kains, PhD
Princess Margaret Cancer Centre,
Department of Medical Biophysics,
University of Toronto, Toronto, Canada

Custos

Prof. Jaakko Kaprio, PhD
Institute for Molecular Medicine Finland (FIMM),
University of Helsinki, Helsinki, Finland

ISBN 978-951-51-4294-8 (paperback)
ISBN 978-951-51-4295-5 (PDF)
Press: Painosalama, Turku

‘We never are definitely right, we can only be sure we are wrong. ‘

- Richard Feynman

Table of Contents

List of original publications

This thesis is based on the following publications:

I. **Jaiswal A,** Peddinti G, Akimov Y, Wennerberg K, Kuznetsov S, Tang J, Aittokallio T (2017) Seed-effect modeling improves the consistency of genome-wide loss-of-function screens and identifies synthetic lethal vulnerabilities in cancer cells. *Genome Medicine* 9 (1):51.

II. Gönen M*, Weir BA*, Cowley GS*, Vazquez F*, Guan Y*, **Jaiswal A***, Karasuyama M*, Uzunangelov V*, Wang T*, Tsherniak A, Howell S, Marbach D, Hoff B, Norman TC, Airola A, Bivol A, Bunte K, Carlin D, Chopra S, Deran A, Ellrott K, Gopalacharyulu P, Graim K, Kaski S, Khan SA, Newton Y, Ng S, Pahikkala T, Paull E, Sokolov A, Tang H, Tang J, Wennerberg K, Xie Y, Zhan X, Zhu F, Aittokallio T, Mamitsuka H, Stuart JM, Boehm JS, Root DE, Xiao G, Stolovitzky G, Hahn WC, Margolin AA (2017) A community challenge for inferring genetic predictors of gene essentialities through analysis of a functional screen of cancer cell lines. *Cell Systems* S2405-4712(17)30392-7.

III. Najumudeen AK, **Jaiswal A***, Lectez B*, Oetken-Lindholm C, Guzman C, Siljamaki E, Posada IMD, Lacey E, Aittokallio T, Abankwa D (2016) Cancer stem cell drugs target K-ras signaling in a stemness context. *Oncogene* 35 (40):5248-5262.

* Equal contribution

The publications are referred to in the text by their roman numerals.

The articles have been reprinted with permission from the copyright holders.

Publications related to the study but not included in thesis

- Kangaspeska S*, Hultsch S*, **Jaiswal A,** Edgren H, Mpindi J-P, Eldfors S, Brück O, Aittokallio T, Kallioniemi O (2016) Systematic drug screening reveals specific vulnerabilities and co-resistance patterns in endocrine-resistant breast cancer. *BMC Cancer* 16 (1):378.

Author contributions

I. Designed the study. Performed the analyses of all datasets. Prepared the figures and wrote the manuscript.

II. Led the analysis team for the winning method in sub-challenge 3. Designed and performed the prediction modelling. Contributed to post-hoc analysis and wrote the manuscript related to sub-challenge 3. Shared first authors led the other teams for the sub-challenge 1 and 2, or were part of the challenge organizers that collected and shared the data.

III. Designed and performed the computational analyses of gene expression and drug sensitivity datasets presented in the study. Prepared the figures and wrote the manuscript related to the computational analyses. The first author was responsible for the experimental data.

Abbreviations

AGE	average gene essentiality
ALK	Anaplastic lymphoma receptor tyrosine kinase
ATARiS	Analytic Technique for Assessment of RNAi by Similarity
BEMKL	Bayesian Efficient Multiple Kernel Learning
BFG	Breast Functional Genomics
BRAF	B-rapidly accelerated fibrosarcoma serine/threonine kinase
bp	base pairs
CCLE	Cancer Cell Line Encyclopedia
CGP	Cancer Genome Project
CNV	copy number variations
CRISPR	clustered regularly interspaced short palindromic repeats
crRNA	CRISPR-derived RNA
CSC	cancer stem cell
CTRP	Cancer Therapeutic Response Portal
DNA	deoxyribonucleic acid
DREAM	Dialogue for Reverse Engineering Assessments and Methods
DRIVE	deep RNAi interrogation of viability effects in cancer
dsRNA	double-stranded RNA
EMT	epithelial-to-mesenchymal
EGFR	Epidermal growth factor receptor
GARP	Gene Activity Rank Profile
GDSC	Genomics of Drug sensitivity in Cancer
geneES	gene essentiality score
gespeR	Gene-specific phenotype estimator
HER2	Human epidermal growth factor receptor 2
ICGC	International Cancer Genomics Consortium
LOO-CV	leave-one-out cross-validation
miRNA	microRNA
MSigDB	Molecular Signatures Database
MT-GRLS	Multi-Target Greedy Regularized Least-Squares
NCI	National Cancer Institute
PAM	proto-space adjacent motif

PARADIGM	PAthway Representation and Analysis by Direct Reference on Graphical Models
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha
PKN3	Protein kinase N3
RIGER	RNAi Gene Set Enrichment
RISC	RNA-induces silencing complex
RNA	ribonucleic acid
RNAi	RNA inteference
RSA	Redundant siRNA activity
seedES	seed essentiality score
sgRNA	single guide RNA
shES	shRNA essentiality score
shRNA	short hairpin RNA
siRNA	short interfering RNA
SPS	seed pairing stability
TA	target abundance
TCGA	The Cancer Genome Atlas
TRC	The RNAi Consortium
UTR	untranslated region

Abstract

Cancer is a leading cause of death worldwide and a major public health burden. Technological advances in high-throughput genomic technologies now allow us to extract gene specific measurements at multiple levels, such as mutation, copy number alterations, gene expression to list a few. Genomic profiling of patient tumors have revealed massive heterogeneity in cancer, making it difficult to pin point the driver genes and translate this knowledge for clinical use. Alternatively, functional profiling based on RNA interference and drug sensitivity screens provide complementary information for understanding the functional relevance of genes related to cancer. Such screens can be used to chart the genetic vulnerabilities of cancer cells which can be useful in exploring therapeutic options. However, undesired off-target effects often complicate the interpretation of the results, and the consistency of these screens have been questioned. With the increasing availability of large-scale data on the molecular and functional characteristics of cancer cell lines, computational approaches are required to extract meaningful information from these datasets. Novel computational methods that are able to account for the complex biological mechanisms involved in RNA interference will improve the prediction of genetic vulnerabilities, and augment the discovery of novel biomarkers and targets for personalized treatment of cancer.

In this work, I have developed and applied novel computational approaches for integration of large-scale genomic and functional datasets. Firstly, I developed an approach to remove noise from genome-wide RNAi screens with the aim to increase their consistency. Further, I applied rigorous statistical analyses in multiple datasets to integrate mutational profiles with genome-wide RNAi screen data to predict novel synthetic lethal partners of major cancer driver genes that were experimentally validated by CRISPR/Cas9 knockout assay. Secondly, I explored the question of predictability of genetic dependencies by developing machine learning models using large-scale genomic datasets to reveal insights into gene dependencies that are more predictable, and identified the molecular features that contribute prominently to such predictions. Thirdly, I show the usefulness of performing computational analysis to identify a gene expression signature associated with cancer stemness, which predicts sensitivity of cancer cells to cancer stem cell

inhibitors. Further, I show that the expression signature is useful in identifying patient sub-groups that will most likely benefit from the therapy. Altogether, the methods developed and applied in this work demonstrate clearly the usefulness of computational approaches to data integration in cancer cell line datasets. These findings advance current translational efforts for cancer therapy under the precision medicine paradigm.

1 Introduction

Cancer is a deadly disease which inflicts havoc on the life of the individual diagnosed with it, also making the experience traumatic and emotionally overwhelming for individuals and families gripped by its influence. With 14 million new cancer patients diagnosed yearly and approximately 9 million deaths, cancer is the second leading cause of death worldwide (*1*). Although substantial progress has been made in terms of understanding its causes as well as development of prevention and treatment strategies (*2*), cancer remains a psychological, social and economic burden, and a major global health challenge (*3*).

Cancer is an outcome of abnormal cellular growth, in which normal cells go awry and disobey the regular rules of tissue growth and differentiation that are necessary for maintaining tissue homeostasis, physiology and function. While normal cells behave in a disciplined manner and are programmed to work in unison with each other to guarantee survival of the organism, cancer cells have only one motive: make more copies of themselves (*4*). Although this nature of cancer was clear from early on, little progress had been made in terms of understanding the causes and the process of carcinogenesis. It was the discovery by Varmus and Bishop in 1976 (*5*), showing that genetic alterations in normal cells had the potential to transform them into cancerous cells, which provided the first coherent view that cancer is a genetic disease. From then on began the modern era of cancer biology, and massive strides have been made in gaining a molecular mechanistic understanding of cancer ever since. With this, also came the realization that cancer is dauntingly complex.

In 2000, Hanahan and Weinberg distilled a giant body of scientific literature on the molecular studies of cancer and tumorigenesis into a generalized conceptual framework called ‘the hallmarks of cancer’ (*6, 7*). They overlayed the molecular and biochemical complexities of cancerous cells with the organizing principles of cellular physiology, and proposed a set of rules that underlie the transformation of normal cells to a malignant phenotype. These acquired capabilities of cancer cells: sustained proliferative signalling, resisting cell death, evading growth suppressors, limitless replicative potential, activation of invasion and metastasis, and sustained angiogenesis – served as a coherent template for making sense of the diverse molecular alterations present in cancer cells. They have also been very useful in interpreting the findings from subsequent genomic studies that followed with the onset of genomic revolution, and has also ushered an era of targeted therapy for treatment of cancer patients (*7*).

Post Human Genome Project the field of cancer genomics blossomed, and several large-scale projects were undertaken to systematically survey the frequency of genomic alterations in specific cancer types (*8*). These studies revealed frequent driver mutations of various kinase genes in melanoma, colon and lung cancer (*9-12*). Further, it was observed that several of the frequent kinase driver mutations were correlated with clinical responses to drug inhibition of the kinase activity (*9, 10*). These observations fortified the previous clinical success of the kinase inhibitor, imatinib mesylate, for treatment of chronic myeloid leukaemia (CML) patients having driving mutations in the BCR-ABL fusion gene, thus setting the stage for arrival of targeted cancer therapy (*13, 14*). The targeted therapy approach requires the identification of molecular targets crucial for the survival of cancer cells in a given genetic background, whose inhibition by a small molecule is expected to be highly selective to killing cancer cells with fewer side effects (*13*). This approach contrasts with the conventional approach of using chemotherapeutic agents that are relatively non-specific and yield considerable side effects.

Spurred by the promise of targeted therapy began a quest to extensively characterize patient tumours (*15-17*). Big consortium projects such as The Cancer Genome Atlas (TCGA) (*15*) and International Cancer Genome Consortium (ICGC) (*16*) were launched for systematic genomic characterization of many cancer types, and are still ongoing. These massive efforts were aided by the maturation of sequencing technologies and the dawn of massively parallel sequencing (MPS), which made it possible to collect variety of genomic information with the same sequencing platform from a large collection of cancer patients (*8*). For instance, the MPS technology could be used in discovering point mutations, detecting copy number variations, quantifying transcript levels, and also in measuring DNA methylation. These studies were quite successful in discovering new driver genes and genetic alterations that have led to an improved molecular level understanding of the processes involved in cancer (*8, 18*).

Contrary to the expectations based on the early success of inhibiting specific driver kinases, the genomic investigations did not reveal many recurrently mutated driver or druggable cancer genes (*8*). Instead, the sequencing studies made it clear that tumors generally harbor multiple genomically altered events, highlighting the incredibly complex landscape of genomic alterations and massive heterogeneity across cancer types, and even within the same tumor (*18*). Moreover, it became a challenging task to identify the genetic alterations that are relevant to cancer survival and growth, and also the presence of multiple genetic alterations mapping to several molecular processes, made it particularly difficult to pinpoint the druggable targets or pathways (*18*). Thus, the aspirations of targeted therapy are still beyond reach, with significant roadblocks in translating the genomic knowledge into clinically actionable treatment strategies.

To fill the gap in the clinical translatability of the deluge of information obtained from the genomic studies, complementary strategies are needed to functionally characterize the variety of genes that are altered in cancer, so as to identify the ones relevant for cancer treatment (*19-21*). *In vitro* loss-of-function screens based on gene suppression using RNA interference (RNAi), or gene inactivation using the recently developed clustered regularly interspaced short palindromic repeats (CRISPR)–Cas9 system have become widely-used techniques for interrogating the role of genes essential in various cancer types (*21, 22*). The ease of scalability of these genome perturbation techniques to high-throughput settings have allowed the examination of the functional roles of genes at genome-scale, thus making it possible to survey the gene essentiality landscapes in panels of cancer cells (*23, 24*). These techniques are also well suited for identifying promising drug-targets, because they mimic the desired effect of drug inhibitors, i.e., reduce the activity of the target protein product (*22*). Similarly, cell-based high-throughput drug sensitivity screens have also been developed to functionally assay the response of cancer cells to a library of small molecules, and are routinely being used to identify promising drug candidates and druggable genetic addictions of cancer cells (*25-31*).

Several projects are being undertaken to extensively characterize the genomic and functional landscapes of a diverse panel of cancer cell line models from a wide variety of histological and tumor backgrounds (*27, 28, 30, 32-37*). Since functional profiling and genomic profiling methods provide complementary information on the cancer cells, these datasets are extremely valuable resources for mining the links between the cancer genotype and phenotype. However, unlike the sequencing based-genomic technologies that are known to be quite robust, functional profiling techniques have several pitfalls. For instance, both RNAi and drug screens are known to suffer from off-target effects, and questions have been raised about the consistency and utility of these data for personalized medicine (*38-40*). Furthermore, the 'big data' nature of these datasets requires the application of sophisticated data analysis techniques and computational algorithms to extract knowledge with potential clinical applicability.

The goal of this thesis is to develop and apply computational and analytical methods that can improve the estimation and prediction of genetic dependencies and druggable vulnerabilities in cancer cells. The ultimate objective is to identify genomic biomarkers potentially linked to effective targeted therapy of cancer. A wide variety of methodologies based on predictive machine learning models, unsupervised clustering, survival analysis and statistical methods are applied for the analytical settings considered in this work. These systems medicine approaches are expected to become important for the emerging translational efforts built on the concepts of personalized medicine and precision oncology.

2 Review of the literature

2.1 RNA interference

RNA interference (RNAi) is a phenomenon of RNA mediated gene silencing. It was first observed in *C. elegans* when long double-stranded RNAs (dsRNA) introduced into the organism led to the cleavage of mRNA transcripts with identical sequences (*41*). Following this discovery, RNAi very quickly became a powerful and widely used tool for genetic screens by gene knockdown. Later studies revealed that several types of RNA molecules could also trigger RNAi, such as RNA viruses, transposons and microRNAs (miRNAs) (*42*). Moreover, exogenously introduced chemically synthesized short RNA duplexes; also called short-interfering RNAs (siRNAs), or endogenously expressed hairpin RNAs; also called short-hairpin RNAs (shRNAs), are also capable of inducing gene silencing (*41*). The discovery of the RNAi pathway has led to a fundamental shift in the understanding of how post-transcriptional gene regulation is achieved in eukaryotic systems. RNAi is known to have important biological functions; for instance, RNAi mediated by dsRNAs plays a major role in viral immunity in plants (*41, 43*). In addition, RNAi triggered by miRNAs, endogenously expressed non-coding RNAs, play an important role in regulation of gene expression during animal and plant development (*41, 43*).

Although RNAi was recognized early on as a widespread phenomenon, present in both plants and animals, its application to mammalian systems revealed that long dsRNAs mediated RNAi triggers the activation of cellular immune response, eventually leading to cell death (*44*). Further biochemical investigations on the mechanistic underpinnings of RNAi machinery in different organisms revealed that short duplex siRNAs are capable of inducing gene knockdown in mammalian cells without activating the immune response (*44*). Chemically synthesized siRNAs that are transfected into cultured cells or shRNAs expressed by genomically integrated viral expression cassettes, are processed by an RNase III enzyme, Dicer, to yield duplex siRNA molecules (Figure 1). siRNAs, usually ~21-23 nucleotides long are the effector molecules of RNAi machinery, which ultimately causes target gene suppression by degrading its mRNA (*44*). However, unlike the effector siRNAs derived from shRNAs or synthetic siRNAs; the effector siRNAs derived from miRNAs do not induce mRNA cleavage and rather repress protein translation by binding to the 3' UTR of target mRNA (*42*).

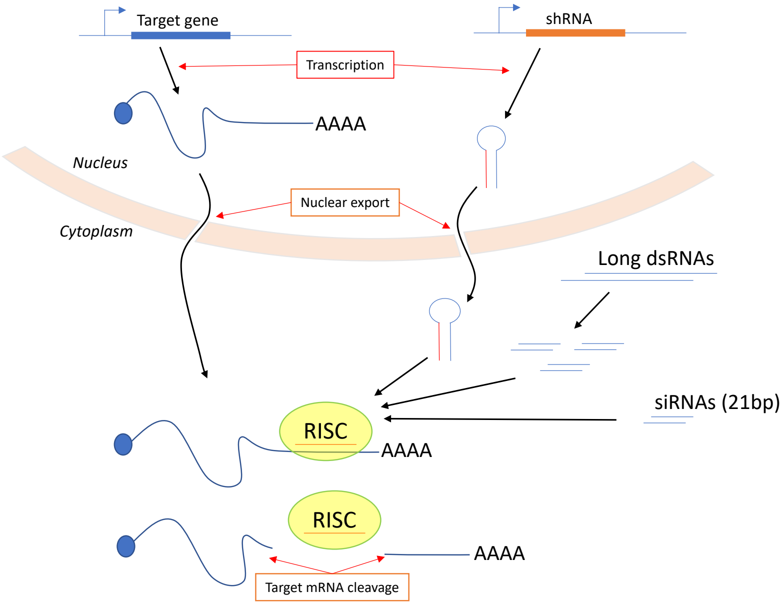


Figure 1: RNAi mechanism of action. Target mRNA and virally tranduced shRNA expression cassettes integrated in the genome are transcribed from their respective promoters. The mRNA product from the shRNA expression cassettes form hairpin structures that are processed further into double-stranded short interfering RNAs (siRNAs). Only one of the strands of the duplex siRNA, known as the 'guide' strand or the 'antisense' strand, is then loaded into a catalytic unit, called RNA-induced silencing complex (RISC). The guide strand serves as a template for guiding the RISC complex to target mRNAs based on sequence complementarity and induce its cleavage in a processive cycle, thereby inhibiting protein translation from the target mRNAs. Adapted from Mohr et al. (45).

2.2 Clustered regularly interspaced short palindromic repeat (CRISPR)/Cas9

CRISPR systems were originally thought to be similar to RNAi (*46*), and were first discovered in *E. coli (47).* Later it was recognized that they play an important role also in immunity to viruses (*48*). When bacteria are exposed to viral or foreign genetic material, short fragments of their DNA are incorporated in the host genome at CRISPR locus separated by a conserved repetitive element (*48, 49*). Transcripts that are generated from a CRISPR locus are processed by CRISPR-associated (Cas9) nucleases into short CRISPR-derived RNAs (crRNAs) that are complementary to the previously exposed foreign DNA material. The crRNAs assemble with Cas proteins to form large complexes that functions as an adaptive immune system in the bacteria, sensing and cleaving any foreign genetic material in the intracellular environment. Later it was realized that the sequence specificity of the crRNA/Cas9 ribonucleoprotein complexes and the ability of Cas protein to create double strand breaks in the DNA can be exploited to conduct genetic perturbations of human cells (*49*). Cas9 can be targeted to specific genomic loci using a 'guide' RNA, which recognizes the target DNA and is able to induce mutagenesis by DNA double-strand break repair pathway. Short single guide RNA (sgRNA) that is complementary to the target DNA is often being used to target the Cas9 nuclease to a desired location in the genome. A sgRNA is typically 20-bp in length and also contains a 3-bp proto-spacer adjacent motif (PAM) after the 20bp region. The cleavage of target DNA, typically a coding region of gene, is induced by the Cas9 nuclease, and loss-of-functions or indels are introduced by the non-homologous end-joining mediated double stranded break repair pathway, creating a knockout of the targeted gene (*49*).

2.3 Genome-wide RNAi screens

The ability to ectopically introduce RNAi agents into cells, and the ease of scalability of the technique to high-throughput settings, has made it possible to perform high-throughput loss-of-function screens, radically enhancing the utility of RNAi to explore a variety of research questions (*50, 51*). Post human genome project era, the availability of complete human genome sequence has allowed the designing of libraries of RNAi agents to conduct genome-wide RNAi screens in human cultured cells and cancer cell lines (*24, 52*). RNAi screens can be performed in human cells either using synthetic siRNAs that are introduced by transfection, or using shRNAs that are expressed from vectors integrated into the host-cell genome. However, the issue of transient gene silencing due to short life of siRNAs inside the cell, the difficulty of efficient transfectional delivery especially to primary cells, and the expensive cost of chemically synthesizing siRNAs, has limited their use in genome-scale screens (*52*).

shRNA-based screens circumvent these problems by using expression cassette vectors that can be stably integrated into the host-cell genome of various cell types using lentiviral or retroviral transduction (*44*). These expression cassettes have promoters that drive the synthesis of shRNA molecules, forming a hairpin structure with 19–29 bp (stem connected by a 6-9 bases long loop, that are processed to generate effector siRNAs (*44*). This provides a stable and renewable source of siRNAs making it possible to study the phenotypic effects of prolonged periods of gene suppression. Moreover, shRNA vectors are amenable to pooled ’barcode’ screens that are less labor intensive, cheap and easily scalable in comparison to plate-based array screens (*52, 53*). Several shRNA libraries are available commercially with varying coverages of the number of genes that can be screened (*24, 52*). For example, The RNAi Consortium (TRC) library covers ~80% of the human coding genes, with an average of six unique shRNA clones per gene. The clones consist of hairpin sequences that are designed based on sequence composition, specificity, and position scoring to increase the likelihood of target gene knockdown (*24*). A desired feature of shRNA library is to have high redundancy in number of clones per gene, which is important in order to reduce false positive results that are due to off-target effects (*24, 52*).

Brummelkamp et al. (53) introduced the idea of using vector encoded shRNA template sequence as a molecular tag or barcodes to quantitatively estimate the abundance of each shRNA vector in the population of cells transduced with a library of shRNA expression vectors. The relative abundance of each barcode sequence can be quantified by PCR amplification coupled with microarray hybridization or next generation sequencing (54). Genome-wide shRNA screens (Figure 2) have been routinely applied to study gene dependency profile of cancer cell lines and identify potential drug targets for cancer treatment (21, 22, 55-57).

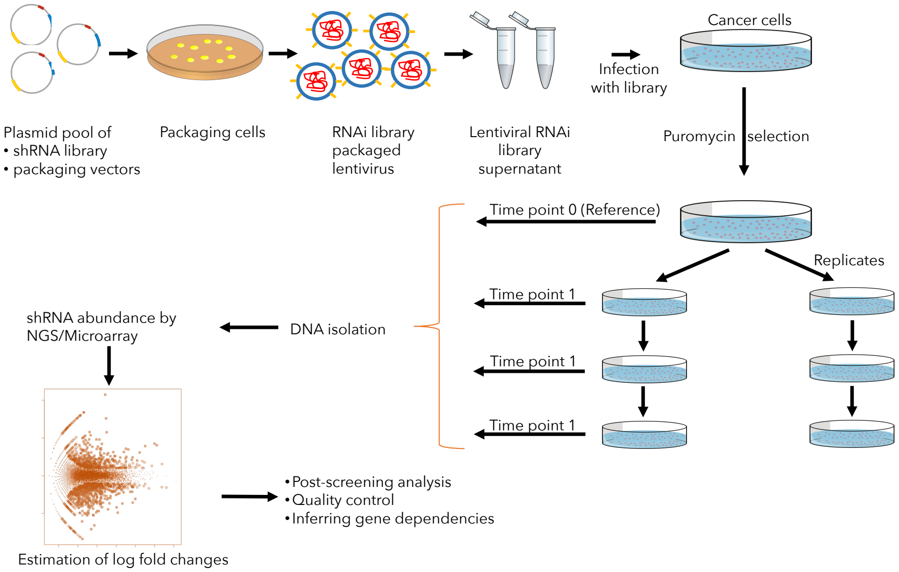


Figure 2: Genome-wide shRNA screen workflow. A pooled genome-wide shRNA screen involves a library construction of pooled plasmids from bacterial culture, followed by viral packaging of the shRNA clones, which is done by transfecting large number of packaging cells together with packaging plasmids. Virus titres produced after 48-72 hours post transfection is pooled and then cell lines of interest for loss-of-function screen are infected and selected to eliminate the uninfected ones. Typically, after this step, an aliquot of the cells is separated and genomic DNA is isolated and used for a quantification of the initial shRNA abundance; and then depending on the experimental design, the cell cultures are divided into two or more sets. For instance, in a drug response modifier screen, cells are divided into treated and untreated aliquots. Alternatively, in a cell viability screen, a sample of cells can be taken and stored for analysis at each passage, to generate a viability time-course. The abundance of each shRNA vector at the final time points is then measured and compared to the initial conditions to get a quantitative estimate of the effect of each shRNAs knockdown on the proliferative capacity of cells (54).

2.4 Off-target effects in RNAi screens

One of the pitfalls of RNAi screening technique has been its propensity to cause off-target effects; therefore limiting its promise and potential (*38, 45, 58*). Transcriptional profiling after inhibition with multiple siRNAs targeting the same genes revealed that the siRNAs also produce strong downregulation of genes other than their primary targets, and moreover each individual siRNA produced a unique fingerprint of transcriptional changes of multiple targets (*59*). Sequence analysis has revealed that 5' end of the guide strand of the siRNAs may have partial complementarity to off-target transcripts, suggesting a sequence-dependent off-target effects (*60*). Biochemical studies also confirmed that 5' end of the guide strand contributes maximally to the target binding and its subsequent cleavage (*61*). Sequence alignment studies revealed that the 'seed' region, which stretches from 2-8 nucleotide positions at the 5' end of the antisense or guide strand of the siRNA, was enriched in the 3' UTR region of the off-targeted transcripts, suggesting a microRNA-like gene silencing pattern (*62, 63*). Alterations in the seed-region of a siRNA or shRNA may also alter the profile of off-targeted transcripts, indicating the importance of its role in mediating the off-target effects (*63*). Likewise, Anderson *et al.* found that siRNAs that have higher number of seed matches to 3' UTR in the transcriptome have a higher propensity towards off-target effects, based on the induced gene expression changes (*64*).

Silencing of off-target genes mainly arises due to the similarity of the siRNA pathway with the endogenous microRNA (miRNA) pathway (*58*). The externally introduced siRNAs utilize and recruit the same components of the downstream RNAi machinery to repress the targets, which is also utilized for normal gene regulation by the miRNAs (*58*). Once the guide strand of siRNA is loaded into the RISC complex and bound to the target, the Argonaute protein of RISC cleaves the target mRNA. Argonaute requires perfect sequence complementarity with the target site to induce cleavage; hence siRNAs can strongly reduce gene expression. In contrast, in the miRNA pathway, complete sequence complementarity of miRNAs with target mRNA is not necessary, and the RISC does not induce target mRNA cleavage (*65*). Thus, miRNA induced gene-silencing leads to translational repression and is incomplete as compared to siRNA induced gene silencing. The partial sequence complementarity in the miRNA pathway is mediated by the seed region, extending from 2-8 nt of the 5' end of the guide strand of the microRNA (*65*). Because of this partial sequence similarity requirement, microRNAs are known to have larger number of target sites that are generally located in the 3' UTR regions of transcripts, and it is estimated that each miRNA may have potentially ~300 target sites (*66*). In addition to target site abundance, other properties such as strength of seed pairing at the target site, its location and spacing in the 3' UTR, local sequence and structural context, are other determinants of miRNA targeting efficiency (*67*). Given the similarity between the miRNA pathway and siRNA pathway, these determinants are also likely to influence the off-target propensity of siRNAs.

Off-target effects have also challenged the interpretability of high-throughput RNAi screens (*68*), with several studies reporting the top hits being false positives. For instance, in a screen designed to identify regulators of HIF1-α transcription pathway, the top siRNAs targeting other genes were still shown to downregulate HIF-1α by an off-target effect mediated through the seed region (*60*). Similarly, in a screen designed to identify modulators of resistance to apoptotic inhibitor ABT-737, the top hits were shown to downregulate another key anti-apoptotic protein, MCL-1, through seed mediated off-target effects (*69*). Sigoillot et al. also observed nonspecific targeting of MAD2 by the active siRNAs in a screen for genes required in spindle assembly checkpoint formation (*70*). These observations highlighted caution in interpreting results from large-scale RNAi screens, and also incited alternate strategies to mitigate the false positive hits (*38, 58*). Using multiple siRNAs per gene, appropriate controls, internal validation with alternative techniques, and performing rescue experiments by expressing a functional version of the target gene, are some of the ways to counter off-target effects in RNAi screens. The false discovery rates in RNAi screens have been discussed extensively (*38*). Meta-analysis of three genome-scale siRNA screens studying host-factors necessary for HIV replication identified virtually no common hits, with <7% overlap between any two screens (*71*). Some studies have also shown that the top hits from a genome-wide shRNA screen for synthetic lethal partners of the oncogene KRAS was not found to be essential in KRAS dependent cancer cell lines, and also did not show any response towards its targeted inhibition (*72-74*). Although the low rate of validation of hits can be due to several factors, such as differences in library, experimental protocols or screened cell lines, and functional redundancy of genes, these observations have raised concerns about the usefulness of large-scale RNAi screens and the reliability of the findings (*39, 75, 76*).

2.5 Methods for inferring gene dependencies from RNAi screens

Genome-scale RNAi screens are experimental techniques that generate massive amount of data, and simultaneously create new challenges for statistical analyses and interpretation to extract meaningful information (*77*). Statistical handling and analysis of RNAi screening data can contribute substantially to the identification of true hits that can influence the consistency and reproducibility of these methods (*77*). The primary goal of a genome-wide RNAi screen is to provide a quantitative estimate of the phenotypic effect specific to each gene in a given cellular context. Computational methods that can take into the account the library design, controls and off-target effects, offer the potential to provide accurate estimates of the gene-specific phenotypes. Several computational methods for estimation of gene dependency scores have been developed, ranging from simple statistical techniques to more sophisticated models incorporating seed-mediated off-target effects of the shRNAs (described below).

2.5.1 Redundant siRNA activity (RSA)

The Redundant siRNA Activity (RSA) analysis method (*78*) makes use of the redundancies in the number of RNAi reagents tested per gene in genome-scale screens to estimate the probability of a gene being a hit. Simply put, the RSA ranks the shRNAs according to their observed quantitative effect and calculates an enrichment p-value based on an iterative hypergeometric distribution method (*79*), similar to pathway analyses based on Fisher's exact text. The p-value indicates the probability of the shRNAs for the gene being distributed towards the top ranks more likely than expected by chance. Because RSA uses probablistic models to infer gene-level phenotypes, it is a powerful approach and outperforms the cutoff based approach of hit calling based on activity of shRNA scores.

2.5.2 RNAi Gene Set Enrichment (RIGER)

RIGER is a non-parametric method (*80*), which shares similarities with the Gene Set Enrichment Analysis (GSEA) technique (*81*) used in differential expression pathway analysis. RIGER utilizes the power of multiple shRNAs per screen to estimate whether they are randomly distributed towards the top or the bottom of the hit list. RIGER calculates gene-level enrichment scores by ranking the entire list of shRNAs, and calculates a running-sum test statistic similar to using a Kolmogorov-Smirnov statistic. Normalized gene-level enrichment scores are then calculated, which takes into account the variability of the number of shRNAs per each gene. The RIGER method does not require any arbitrary threshold to estimate the enrichment scores. Directional RIGER (dRIGER) (*82*), an extension of RIGER, has also been used for transforming shRNA-level scores into gene-level scores by computing directional normalized enrichment scores (dNES).

2.5.3 Gene Activity Rank Profile (GARP)

GARP score (*83*) takes into account the dropout behaviour of the shRNAs across several time points. First, a summarized shRNA activity ranking profile (shRNA) score is calculated by averaging the relative change in shRNA abundances, which is normalized by the number of population doublings in the assay. Then, from the multiple sets of shRNAs targeting the same gene, the average of two shRNAs with lowest shARP scores is considered as the GARP score. Statistical *p*-values are calculated from permutation testing across 1000 random scores, as a measure of the statistical 'significance' of an observed GARP score.

2.5.4 Analytic Technique for Assessment of RNAi by Similarity (ATARiS)

ATARiS (*84*) evaluates the quantitative behaviour of shRNAs targeting the same gene across various samples to identify the shRNAs that are likely to produce on-target effects. For identifying the on-target shRNAs, ATARiS creates a consensus profile from the activity profiles of all the shRNAs against a gene in several samples by using information divergence and alternative minimization techniques, which separates the shRNA-specific effects from the consensus effect. Then, the algorithm performs iterative correlation analysis of each of the shRNAs with the consensus profile, and discards the ones that are statistically insignificant and recomputes the consensus profile. The final consensus profile based on the on-target shRNAs is used as the gene-level score. Further, the algorithm also calculates a consistency score for each shRNA reagent, indicating the likelihood of its on-target effect. Because ATARiS considers the consistency of shRNA effects across several samples, the number of samples used in the analysis also influence the number of genes for which the final scores are derived.

2.5.5 Gene-specific phenotype estimator (gespeR)

gespeR (*85*) performs a statistical modelling for the estimation of gene level scores by taking into account the on-target and off-target activity of the shRNAs. gespeR uses elastic net regularization to fit a linear regression model on the observed shRNA activities against a shRNA-target gene relationship matrix. The shRNA-target gene relationship matrix is obtained by using the TargetScan algorithm (*67, 86*), which quantitatively predicts the probability of knockdown of off-target genes for each shRNA based on its seed sequence. TargetScan also considers other properties of shRNA sequences, such as seed pairing stability, target abundance and 3' UTR location of target site and local AU context to predict the knockdown efficiency of off-target genes. The final regression coefficients derived after cross-validation are considered as the gene-level scores.

2.5.6 DEMETER

DEMETER (*87*) assumes that each shRNAs phenotypic effect is a linear combination of target gene knockdown effects and seed-specific effects. DEMETER takes into account the numbers of shRNAs per each gene in the library, and also the numbers of shRNAs with the same seed sequence. For each shRNA, it considers two seed sequences positions, 1-7 and 2-8 of the guide strand. DEMETER performs deconvolution of the shRNA level data into a linear combination of gene and seed-level effects using stochastic gradient descent. It also provides a performance metric for each shRNA, a measure of the variance explained by gene effect and seed effect. It was recently shown that the removal of seed effects from shRNA level data led to a substantial improvement in the correlation of shRNAs targeting the same gene (*36*).

2.6 Functional genomic characteristics of cancer cell lines

Large-scale sequencing efforts, such as TCGA and ICGC, have aided massively in our understanding of the major genetic alterations in cancer genomes, in addition to providing an overview of the genomic landscapes. The cancer sequencing studies have catalogued an impressive list of new genes, previously unknown to be involved in cancer with some genes more frequently mutated than others. While these studies are ongoing and identifying more genes associated with cancer, alternative strategies are also required to make a sense of the plethora of genetic alterations. Loss-of-function screens based on RNAi and CRISPR/Cas9 are suitable methods for understanding the functional implications of the cancer-associated genes, which can lead to a better understanding of the dependencies of cancer cells on certain genes or biological processes. Several efforts are being carried out to functionally characterize large collections of cancer cell lines with genome-wide loss-of-function screens, along with characterizing their genomic features including mutations, copy number variations, transcriptome, proteome and the epigenomic profiles. Integrated analysis of these datasets can provide valuable insights about the biology of cancer, as well as identify biomarkers for patient stratification for the right treatment strategy and novel targets for targeted anticancer treatment.

2.6.1 Cancer cell lines as models for anticancer therapies

Preclinical models, such as human cancer-derived cell lines, have contributed immeasurably to the understanding of the biology of cancer (*88*). The advantages of *in vitro* cancer cell lines are multifold: they can be easily cultured, are renewable, are amenable to high-throughput assays, can be easily adapted to sophisticated experimental designs like studying drug resistance modulators, or response to combinations of drugs. Moreover, linking the molecular and genetic features of cancer cell lines with their phenotypic and drug sensitivity profiles has the potential to identify promising biomarkers for targeted therapy (*89*). The National Cancer Institute (NCI) resource (NCI-60), that characterized a panel of 60 cancer cell line models representing 9 different cancer types was the first cell line resource initially setup to screen the activity of a large library of compounds (*89-91*). Initial studies revealed that drugs with similar drug response profiles were similar in their mechanism of action, suggesting that cellular state influences the phenotypic responses (*92*). More importantly, studies of drug response profiles in NCI-60 panel led to the identification of the proteasomal inhibitor, bortezomib, for treatment of patients with multiple myeloma, hence highlighting the usefulness of the cell line based functional screens (*93, 94*). Later, it was also found that gene expression features are correlated with drug responses, suggesting that molecular features of cell lines can be used to predict their functional phenotypes (*95*).

Genomic characterizations of NCI-60 and other cancer cell line panels have revealed that they retain the recurrent genetic and epigenetic alterations present in tumors (*92*). Moreover, cancer cell line models also mimic their sensitivity to targeted drugs, for example, lung cancer cell lines with oncogenic driver alterations, such as EGFR, BRAF mutations, ALK translocations and HER2 amplifications, retain their sensitivity to the respective kinase inhibitors, suggesting that they also able to recapitulate the therapeutic response profile of tumors (*88, 96, 97*). However, contradicting observations have been made for the comparisons at the transcriptome level (*98*). Lukk et al. performed a combined analysis of gene expression data of cancer cell lines and patient tumors representing similar tissue types, and observed that the cancer cell lines clustered together with each other rather than with the tumor samples of the respective tissue type (*99*). In contrast, Ross et al. observed that breast cancer cell lines were able to faithfully recapitulate the tumor subtypes based on the gene expression data (*100*). Additionally, Barretina et al.

demonstrated that huge compendiums of cancer cell lines mirrored the architecture of human tumors suggesting that profiling a larger panel of cancer cell lines would be required to recapitulate the heterogeneity present in patient tumors (*27*). Based on the genomic studies on patient tumors by TCGA and other consortia, it was realized that more cell lines need to be profiled to capture the genetic variability (*36, 89*). Hence, several projects have been undertaken to molecularly and functionally characterize larger panels of cancer cell line models to recapitulate the heterogeneity associated with patient tumors (*27-29, 33, 35-37, 101-103*). The use of cancer cell lines for drug discovery efforts have also been questioned (*104*). As they are grown *in vitro* on plastic surfaces, they do not recapitulate the tumor microenvironment and the drug pharmacokinetics. Moreover, it has been observed that the adaptation of cells to the plastic surface introduces new mutations and genetic aberrations that might change their genetic characteristics (*105-107*).

2.6.2 Genomic profiling of cancer cell lines

To model the genetic diversity of tumors, several large scale, pan-cancer efforts such as the Cancer Cell Line Encyclopedia (CCLE) (*27*), Cancer Genome Project (CGP), and its resource called Genomics of Drug sensitivity in cancer (GDSC) (*28, 29, 108*), and Genentech Resource (*109*) have recently been undertaken to molecularly characterize panels of cell lines from various tumor types. Tissue-type specific panels such as breast (*110*), ovarian (*90*), non-small lung cancer, head and neck cancer (*111*) and colorectal cancer (*112*) cell lines have also been profiled separately. Comparison of copy number variations (CNV) and gene expression profiles of breast cancer cell lines with tumors established that the functionally important alterations were preserved, with 72% agreement of the gene expression changes (*110*). Interestingly, a greater number of CNVs were observed in the breast cell lines underscoring the caution in clinical interpretability of observations from cell lines (*110*).

Cancer cell lines from several solid tumor types, including ovarian, head and neck and colorectal cancer, closely resemble the mutational profiles of their respective tumors, but have higher number of point mutations (*111-113*). Whereas the CNV profiles of head and neck cancer cell lines were different from the tumor samples (*111*), good agreement of the CNV profiles of colorectal (*112*), melanoma (*114*), non-small cell lung cancer (*115*) was observed. A large panel of cell lines characterized by CCLE, approximately 1,000 cell lines representing 36 cancer types, also showed strong correlation of all three genomic profiles: mutation, CNVs and gene expression with their respective tumor types in most cases (*27*). In the same vein, the GDSC project, which profiled ~1000 cell lines representing 29 tumor types, also revealed good agreement in the mutational landscapes (*28*). The GDSC study observed high levels of agreement between functional events that were defined as clinically relevant, with 1063 present in cancer cell lines out of 1273 events present in tumors (*28*). In addition, the authors also reported high agreement for pathway level alterations and global signatures of events associated with driver mutations.

Transcriptomic analysis of 675 cancer cell lines comprising of 18 tissue types from the Genentech resource revealed that the lymphoid cell lines clustered separately from the set of cell lines or other tissue types as observed in previous studies (*99, 109*). Moreover, the latter group further sub-clustered into epithelial and mesenchymal subtypes correlating with the classification based on genes associated with epithelial-to-mesenchymal transition (EMT)-signature (*116*). Although EMT is a transdifferentiation program activated in cells during embryonic development (*117, 118*), its induction has also been correlated with invasive and metastatic potential of cancer cells during tumor progression (*116, 119-123*), and more importantly with the emergence of drug resistance (*124-126*). Importantly, the acquisition of mesenchymal traits through EMT is associated with the expression of stem cell markers, i.e. a cancer stem cell (CSC)-like phenotype (*119, 127*). CSCs are known to self-renew and contribute to tumor heterogeneity and are resistant to chemo- and radiation therapy (*126, 128-130*). Several studies have identified subpopulations of CSC-like cells in cancer cell lines from breast (*131-134*), glioma (*135*) and head and neck cancer (*136*), demonstrating that cancer cell lines can also be used to study the survival mechanisms of CSCs.

2.6.3 Functional profiling of cancer cell lines

Lessons from genomic studies of cancer cell lines have fortified their use as faithful models for expediting the discovery of effective targets for precision anticancer treatment. However, these studies do not provide answers on whether the identified genomic alterations are important for the tumor biology, and whether they yield a therapeutic opportunity as druggable targets. Hence, several large-scale efforts based on loss-of-function and drug sensitivity screens have also been undertaken to functionally characterize the cancer cell line panels. Project Achilles (*32,* *36, 102*), by the Broad Institute, performed systematic genome-wide RNAi screen of 501 cancer cell lines, representing 30 different cancer types and identified ~750 genes that are differential essential in cancer cell lines (*36*). The authors observed that only 76 genes from this set was present in almost 90% of the cell lines, suggesting that the same essential genes are relevant across many tumors. Moreover, a substantial proportion of the essential genes were also druggable (*36*).

An earlier report from Project Achilles also revealed essential genes that are tissue-specific and aberrantly activated due to amplification or overexpression in multiple cancer types (*102*). The Project DRIVE also interrogated the functional effect on cell viability of ~8000 genes by genome-wide shRNA library in nearly 400 cancer cell lines, representing 26 cancer types and identified the dependence of cancer cell lines on lineage-specific transcription factors (*33*). Marcotte et al. observed that the gene essentiality profiles of breast cancer cell lines partially corresponded to the breast tumor subtypes, in addition to observing driver mutation-specific and cancer type-specific dependencies (*83*). The COLT-cancer database comprises of functional profiles from genome-wide shRNA screening of ~15000 genes in 72 cancer cell lines from pancreatic, ovarian and breast cancer types (*37, 83*). In another study on a larger panel of breast cancer cell lines, Marcotte et al. identified gene dependencies in EGFR and MAPK pathway genes that were correlated with the response of the cell lines to targeted inhibitors of EGFR/MEK/ERK (*34*). Recently, genome-wide CRISPR/Cas9 based knockout screens have also been performed in large panel of cancer cell lines (*137-141*), revealing potential targets for acute myeloid leukemia (*139*), and vulnerabilities important in the context of KRAS mutated cancer cells (*137*).

In addition to the functional profiles based on loss-of-function screens, several studies have performed drug sensitivity profiling of cell lines against a library of small molecules. The CCLE profiled the activity of 24 targeted and cytotoxic agents against cancer cell lines at several doses, and by performing predictive modelling with elastic-net regression, they identified several genomic predictors of the drug responses (*27*). Similarly, the Cancer Therapeutic Response Portal (CTRP) (*30, 31*) and GDSC (*28, 29*) projects have also profiled the activity of a library of drugs, 480 and 265 respectively, in a larger panel of cell lines. Drug sensitivity screens have also been used to identify CSC-specific inhibitors in breast epithelial cell lines induced to undergo EMT (*142*). Although drug sensitivity screening is not a functional genomics tool in its true sense, it provides complementary information on the phenotypic characteristics of the cell lines, and has led to identification of novel drugs for cancer treatment (*26, 143, 144*). However, drug screens also suffer from the off-target effects and promiscuity of inhibitors to modulate related proteins, making it difficult to attribute the observed drug responses to their primary targets, also called target deconvolution problem of phenotype-based drug discovery approach.

2.6.4 Consistency of functional and genomic datasets

With the availability of genomic and functional profiles of cancer cell lines from different laboratories, a natural question that arises is how consistent these profiles are. Cancer cell lines are known to acquire genetic aberrations during the culturing process, and because cancer cell lines are widely used across research labs, it is important to understand whether the datasets generated from the panels of cell lines by various studies draw a consistent portrait. In addition, the consistency of the datasets can also be influenced by laboratory protocols and workflow, experimental factors such as cell confluency, genomic drift, clonal variations, growth medium, the robustness of the platform being used for high-throughput measurement and computational methods used in data post-processing (*145*).

Genomic platforms are known to be quite robust and extensive work has gone into standardizing workflows and data processing pipelines. Encouragingly, comparison of the transcriptomes of cell lines profiled commonly in the Genentech Resource with CCLE and CGP have revealed nearly 80% agreement between the datasets (*109*). Comparison of gene expression and mutational profiles between CCLE and CGP also indicated high correlation levels (*109*). In contrast, the consistency of drug sensitivity screens has been a matter of recent debate with several groups reporting dissimilar observations (*40, 146-148*). Originally, Haibe-Kains et al. observed only ~30% agreement between drug responses measured in CCLE and CGP (*40*). In subsequent analysis, Mpindi et al. observed that correlation of the profiles could be increased up to 70% by using standardized metrics of quantifying drug sensitivity, and by standardizing assay methods and protocols (*148*). It was also observed that higher concordance can be achieved by using more biologically motivated statistical analysis methods, and accounting for experimental factors like cell seed density and cell growth media (*149*). Functional profiles based on genome-wide RNAi screens are also known to be noisy and inconsistent, mainly due the off-target effects mediated by partial complementarity (*38, 58*). However, systematic comparisons of the consistency of RNAi or CRISPR/Cas9 datasets have not been performed.

2.7 Integrating genomic and functional profiles

The goal of precision medicine and targeted cancer therapy is to identify biomarkers that will help tailor the best treatment option for each patient. Treatment of breast cancer patients overexpressing HER2 receptor with HER2 antibodies, and leukemia patients harboring BCR-ABL fusions with imatinib are some successful examples, based on the idea of oncogenic addiction, demonstrating how single genomic markers can guide effective cancer treatment (*8, 18*). However, the genetic alterations in many cancer driver genes do not always correspond to it being essential for survival. Extensive genetic heterogeneity resulting from multiple alterations also makes it difficult to pinpoint the specific dependencies in cancer cells. Integrative analysis of molecular features of cancer cell lines and their functional profiles can be used to identify the genetic dependencies associated with a certain genetic background.

2.7.1 Beyond oncogene addictions: synthetic lethality

Synthetic lethality is defined as the significant reduction of cellular viability due to simultaneous loss-of-function of two partner genes, such that when the genes are inhibited individually they do not compromise the cell viability (*150-153*). Cellular signalling is a robust process with several feedback loops and functional redundancies which ensure that cells are capable of surviving when a certain genes' function is lost or inhibited (*151*). Thus, simultaneously inhibiting these functionally redundant genes to compromise the viability of cancer cells is a promising strategy for anticancer treatment (*152, 153*). The idea is to exploit on the vulnerability of cancer cells; having a frequently occurring genomic alteration makes the cancer cells more dependent on the synthetic lethal partner for survival. It is expected that only cancer cells harbouring the genetic alteration will be sensitive towards the inhibition of the activity of the synthetic lethal partner gene, hence having a broader therapeutic window and less side effects in normal cells (*154*). Moreover, targeting synthetic lethal partner of tumor suppressors, which already have loss-of-function mutations, is especially beneficial as they are not easily amenable to drug inhibition (*154*). The synthetic lethality approach is different from the concept of 'oncogene addiction' which is based on inhibiting the activity of single altered driver oncogene, such as HER2, BCR-ABL, EGFR and BRAF (*154*).

Synthetic lethality provides also a framework for associating the genomic features of cancer cells with their phenotypic characteristics. Functional profiles from genome-wide loss-of-function screens in cancer cell lines are a rich source of information for identifying novel synthetic lethal interactions and have been used routinely in the past (*22, 155*). Frequently occurring genetic alterations of cancer driver genes are associated with changes in the cellular signalling and processes, which renders the cancer cells being vulnerable to their inhibition. For instance, mutations in the BRCA1 and BRCA2 genes are associated with sensitivity of the cancer cells towards inhibition of DNA repair machinery (*156-158*). BRCA genes are involved in repair of DNA breaks by homologous recombination, and thus the inhibition of PARP genes that are involved in base excision repair results in a strong synthetic lethal interaction with BRCA. Several synthetic lethal screens in cancer cell lines have identified putative synthetic lethal partners of undruggable cancer driver genes, such as KRAS, MYC and TP53 (*159*). Genome-wide RNAi screens in panels of mutant KRAS and wild-type cell lines or isogenic cell line pairs identified several synthetic lethal partners such as PLK1, SKT33 (*160*).

However, it has been difficult to translate these findings to a clinical setting due to lack of supporting evidence in other cell lines, *in vivo* models or by drug targeting. So far, only one anticancer treatment based on the synthetic lethal strategy has progressed to the clinical practices, namely, the approval of PARP inhibitors for treatment of breast cancer patients with germline BRCA mutations (*159*). One reason for such disappointing clinical translation rate is that robust synthetic lethal interactions are difficult to identify, as they are known to be highly context-dependent and influenced by the genetic background or microenvironment of the tumors (*161*). Moreover, genome-wide RNAi screens are known to be noisy and contain wide off-target effects, which further make it harder to detect the true synthetic lethal hits from the background noise. It has been argued that integrated analyses to identify robust, context-specific synthetic lethal interactions a panel of cell lines from a variety of lineage backgrounds and various genomic and functional datasets may lead to the identification of clinically actionable synthetic lethal partners of cancer driver genes (*162*).

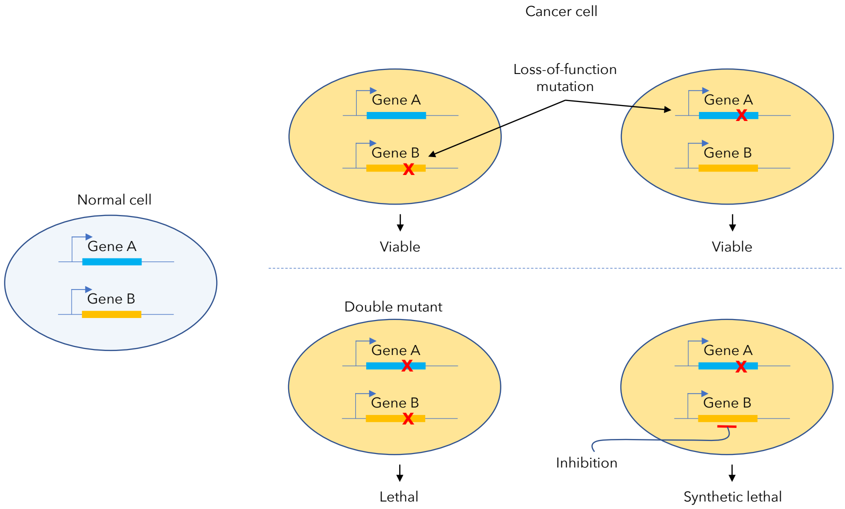


Figure 3: Concept of synthetic lethality for cancer treatment. While normal cells have functional protein products of both gene A and gene B, cancer cells may have loss-of-function mutations in either gene A or gene B individually and are are still viable. However, loss-of-function of both genes in the same cell either by mutation or knockdown or pharmacological inhibition results in synthetic lethality. Modified from O'Nell et al. (161).

2.7.2 Machine learning models for predicting functional profiles in cancer cells

In recent years, the application of machine learning methods in the field of genetics and genomics has tremendously increased and also proved to be very useful (*163*). For instance, machine learning can be used to identify the location of transcription start sites, promoters, splice sites or enhancer sites in the genome (*163*). Artificial intelligence is another field of computer science that deals with the science of making machines that are able to perform intelligent and rational tasks, akin to thinking like humans; machine learning on the other hand is a way of achieving artificial intelligence (*164*). Machine learning involves the building and application of algorithms with the ability to 'learn' i.e. become better at a given task with experience. It is therefore a data-oriented field geared towards problems for which data is available, and on which we can learn and get better at prediction. Machine learning methods can be principally categorized into: supervised learning or unsupervised learning.

Supervised learning requires the use of labels or known examples to train the algorithm, which is then used to predict the respective labels of unlabelled cases. In contrast, unsupervised learning is concerned with finding patterns or clusters in unlabelled data sets, i.e. without any prior knowldege (*163*).

The availability of large-scale functional screening profiles of cancer cell lines that are also molecularly characterized allows the possibility to build computational models that can capable of predicting the phenotypic responses and also identify the relevant genomic biomarkers. Building predictive computational models is a challenging task because of reasons such as: molecular heterogeneity of cancer types, data complexity in terms of size, noise, and standardization and normalization of datasets from multiple sources. However, machine learning algorithms are well-suited for building predictive models (*165*). Several machine learning models, such as support vector machines, elastic net regression, neural networks and random forest, are commonly used to solve such problems (*163*).

The core idea is straightforward: given the genomic features of cell lines and their functional profiles as input, the task is to learn a model that can predict the gene dependencies or drug responses in unseen cell lines or tumor samples. The standard strategy for developing such supervised machine learning models is as follows: first obtain the relevant normalized datasets containing the molecular features that is used as features, and the functional profiles which is considered as the predictor variable. In the second step, the predictive model is trained and selected using the input data using several statistical and machine learning frameworks. The model choice is dependent on the characteristics of the input and output datasets. While nonlinear models can capture complex interections between the input features in the dataset, linear models are easier to interpet, scalable and therefore more preferred (*166*). Thirdly, the trained model is tested on independent datasets to verify the predictive accuracy of the model (*165*).

Supervised machine learning tasks can be categorized into regression and classification problems. In regression models the data to be predicted is a continuous variable, whereas the later deals with variables that are categorical in nature, such as, high response vs. low response. In regression models, the functional profile which is to be predicted can be modelled as a linear combination of the predictor variables.

$$y = \mu + \mathrm{X}\beta + e$$

Where *y* is a vector of the observed functional profile to be predicted, µ is the intercept, X is a matrix of the molecular features, such an gene expression and copy number varation, and β is the regression coefficient and *e* is the vector of residual errors.

Regularized regression models are often used to control model complexity, to avoid overfitting of the training data and enable the generalizability to unseen data. Regularization approaches introduce penalty terms such as L1 and L2 norms for regression coefficients. Ridge regression solves the problem using L2 penalized least squares and is suitable in cases when there are many predictors with small effects. It is formulated as

$$\hat{\beta}(ridge) = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

where,

$$\|y - X\beta\|_2^2 = \sum_{i=1}^{n} (y - x_i^T \beta)^2$$

is the residual sum of squares loss function, and x_i^T is the *i*-th row of *X,*

$$\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$$

is the L2 norm penalty, and $\lambda \geq 0$ is the tuning parameter, also known as the regularization parameter. The regularization parameter is used to shrink the variable coefficents towards zero to prevent any particular variable from having too large effect on the model. In contrast, lasso uses the L1 norm penalty to build sparse models with few non-zero coefficients, and hence suitable for feature selection.

$$\widehat{\beta}\,(lasso) = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

where,

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

is the L1 norm penalty used for introducing sparsity in the solution, with $\lambda \geq 0$. Elastic net regression uses a mixture of the L1 and L2 penalties and can be thought of as an extension of lasso, but with the property to select variables that are still highly correlated.

$$\widehat{\beta}\ (elastic\ net) = \left\{ \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}$$

Systematic analyses of various modelling approaches by Jang et. al. previously showed that the predictive accuracy is determined largely by the type of molecular data used as input and the choice of learning algorithm (*167*). They found that elastic net and ridge regression models applied to continuous response variables, such as drug response, were the best predictors. Moreover, modelling choices are also dependent on their ability to utlize multiple datatypes. For example, gene expression data is a continuous type of input data, whereas copy number variation data and mutation data can be categorical. A commonly used model building strategy is to combine all the input molecular features and train a single model. Other strategies of multiview learning based on multiple kernels have also been applied (*168, 169*).

Machine learning algorithms based on linear models have been generally well formulated theoretically. However, real world problems involve complex relationships between variables that are often not captured by the linear modelling assumptions. Non-linear modelling approaches can be better in detecting these dependencies. To model the non-linear relationships, kernel methods have been used regularly over the past few years. Kernel methods map the data into higher dimensional space using a kernel function such that the data becomes well structured and separated. Essentially, a kernel is a dot product between two feature vectors which is also used as a similarity measure. The advantage with using kernel functions is that when the input data is mapped into a higher-dimensinal space, a linear dependency that exists in this space will behave non-linearly in the original input space. (*170*) Non-linear models have also been used previously to predict drug sensivities in cancer cell lines (*169*). The top performing model in the NCI-DREAM Challenge, which compared the performance of machine learning models for predicting drug responses in breast cancer cell lines among several teams, was based on kernelized regression. The winning algorithm, known as Bayesian multitask multiple kernel learning (MKL) method leveraged four machine-learning principles: kernelized regression, multiview learning, multitask learning, and Bayesian inference. The kernelized regression approach computes kernels, similar to support vector machines, between the cell lines which reduces the number of model parameters and also captures the non-linear relationships between the molecular features.

Several studies have developed predictive machine learning models of drug response in the CCLE and GDSC dataset, using molecular and genomic information available from gene expression, CNV and mutation data (*27-29, 171-175*). The NCI-DREAM Challenge found that the best performing machine learning model integrated information from various datatypes, such as genomic, proteomic and epigenomic profiles (*169*). The study also observed that gene expression data was the most predictive compared to all other data types (*169*). Likewise, Kim et al. developed a computational method based on mutual information metric to identify combinations of mutually exclusive genomic features that are associated with the gene dependency profiles as well as drug response profiles (*176*). These *in silico* prediction tools have been used to predict drug responses in cancer cells, and can potentially help the prioritization of promising drugs and targets for further research and clinical validation leading to significant reduction in experimental costs.

However, a key challenge in building predictive machine learning models is the high dimensionality of genomic datasets and the low sample sizes, also called as the ‘small n, large p’ problem (*177*). This often leads to low power and makes robust inference problematic, and consequently the clinical translatability of these models is often limited. Also, the predictive models are built on a simplistic paradigm: learning a regression model between the predictors and outputs. With the availability of multiple data sources, it is advantageous to model the shared relationships between the variables to increase the predictive performance, using so-called multi-task learning approaches. Another shortcoming of the existing modelling approaches is that they are heavily dependent on ‘statistical’ treatment of the data. Thus, there is a need to integrate the several layers of biological information in a systems modelling framework that takes into account the dynamic cross-talk and network properties of genes (*178*).

3 Aims of the study

The primary aim of this thesis is to develop and apply novel computational methods to integrate the functional and genomic characteristics of cancer cell lines. The goal is to identify promising drug candidates and predictive biomarkers for targeted anticancer treatment strategies. The specific aims can be summarized as follows:

1. Develop and assess computational approaches to increase reproducibility of gene dependencies identified in cancer cell lines based on genome-wide RNAi screens.
2. Predict novel synthetic lethal interactions in cancer cell lines based on normalized RNAi screening datasets, sub-sequently confirmed with CRISPR/Cas9 assays.
3. Develop computational approaches for predicting gene dependencies in cancer cell lines based on their integrated genomic and molecular profiles.
4. Predict drug response of cancer cell lines to cancer stem cell inhibitors using novel transcriptional signatures.

4 Materials and Methods

In this section, I will briefly describe the data sets used in this thesis, along with experimental procedures and cell lines, and the computational models and statistical analysis that were applied in the studies. A detailed description of the materials and methods can be found in the original publications (**I-III**).

4.1 Datasets

Publication	Data set	Data type	Material	Number of samples
I	Project Achilles 2.0 (*102*)	shRNA screen	Cell lines	102
I, II	Project Achilles 2.4 (*32*)	shRNA screen	Cell lines	215
I	COLT-Cancer (*83*)	shRNA screen	Cell lines	72
I	BFG (*34*)	shRNA screen	Cell lines	77
I, II	CCLE (*27*)	Mutation	Cell lines	1074
II, III	CCLE	Expression	Cell lines	1074
II	CCLE	CNV	Cell lines	1074
III	TCGA (*179*)	Expression	Patient tumors	8226
III	ESTOOLs (*180*)	Expression	Embryonic stem cells and fibroblasts	653

Table 1: Data sets and datatypes used in each publication included in the thesis.

4.2 Cell lines for profiling experiments

MCF10A cell line harboring PIK3CA mutations and its corresponding wildtype isogenic controls were purchased from Horizon Discovery Group (city, country) for publication **I**. These cell lines along with 293 packaging cell line for lentiviral packaging was used for CRISPR/Cas9 knockout screening. 15 cancer cell lines used for drug sensitivity testing in publication III were profiled by GenScript profiling services (Finland).

4.3 CRISPR/Cas9 knockout assay

For publication **I**, single guide RNAs (sgRNAs) against target genes were obtained from SigmaAldrich (Helsinki, Finland) and lentiviral particles were generated by transfection using lentiviral plasmids and packaging plasmids. Cas9 expression cell lines were generated and transfected with lentivirus particles packaged with the sgRNAs.

4.4 Statistical analysis

Rank-based Spearman correlation was used for assessing the concordance of essentiality phenotypes (publication **I**), for evaluating the agreement between predicted and observed gene essentiality scores (publication **II**), and for identifying co-expressed genes (publication **III**). Paired t-test and Wilcoxon rank sum tests were used for comparing normal and non-normal distributions, respectively, in publication **I**. Permutation-based statistical testing was carried out in publications **I** and **II** to assess the statistical significance of different types of observed quantities. The advantage of permutation tests is that it does not require any distributional assumptions, and hence useful when the actual distribution is unknown, or when the sample sizes are not large enough for large-sample assumption.

4.5 Survival analysis

TCGA datasets obtained from cancer patients also contain clinical information related to patient survival. Survival analysis related to given a biomarker or gene signature is a useful method for assessing its clinical relevance, in which a comparison of survival time is done between two patient groups, defined based on the biomarker or signature. Kaplan-Meier survival analysis was performed in publications **II** and **III** to assess the effect of the identified gene expression signatures on patient survival in TCGA datasets.

4.6 Clustering analysis

Unsupervised hierarchical clustering was used in publication **III** for assessing the gene expression signature pattern in ESCs and fibroblasts, and cancer cell lines. Clustering methods are useful in resolving patterns and identifying sub-groups in a complex dataset, and routinely used in analysing gene expression profling data. There are several agglomerative clustering approaches, in which the clustering process starts from the bottom. Specifically, each gene or sample is first considered as a singleton cluster, and sub-sequentially these clusters are merged together to form larger clusters, and eventually the entire dataset is part of a big cluster. The sequence of merging each node into larger clusters can be represented as a dendrogram.

4.7 Machine-learning models

In publication **II**, the Multi-Target Greedy Regularized Least-Squares (MT-GRLS) algorithm based on linear modelling was used. MT-GRLS constructs a multi-target ridge regression models given a budget restriction on the number of common features to be selected for performing the multiple tasks i.e. gene essentiality predictions (*181, 182*). For selecting the genomic features, MT-GRLS performs step-wise greedy forward selection, starting with an empty feature set, and then in each iteration adds the feature whose addition results in the maximal accuracy gain, e.g, minimum sum of squared error in the leave-one-out cross-validation (LOO-CV). MT-GRLS optimizes the predictive performance subject to an explicit joint budget constraint on the number of features. The advantage of MT-GRLS algorithm is that it performs the feature selection computationally much more efficiently compared with a straightforward wrapper type of implementation. In a dataset with *d* features and *m* training samples, the time complexity of a standard wrapper approach using LOO-CV for forward selection of k common features for simultaneous prediction of t tasks with RLS would be $O(min\{k^3m^2dt, k^2m^3dt\})$. In contrast, the time complexity of MT-GRLS is only $O(kmdt)$.

4.8 Broad-DREAM Gene Essentiality Prediction Challenge

The DREAM Challenges are crowdsourcing challenges examining challenging questions in biology and medicine (*183*). The DREAM challenge organizers pre-test all data and predictions, and develop custom scoring methodologies to ensure high-quality data and rigorous performance evaluation. Such crowdsourcing competitions produce standardized data sets and benchmarked methods for future comparison, analysis, and model development (*184*). Crowdsourcing competitons can be a useful approach to doing scientific research. It can reveal a variety of approaches towards solving the same task. It can also reveal biases in scientific conclusions that are based on subjective analytical approaches. Since each pariticipating team contributes to the competition with its own findings, a range of results are revealed which can be useful in guiding the research towards a fruitful direction (*185*).

The goal of the Broad-DREAM Gene Essentiality Prediction Challenge was to use a crowd-based competition to develop predictive computational models that can infer gene dependencies of cancer cells using their molecular and genomic features. Participants were provided with gene essentiality datasets generated from genome-scale shRNA screen in a panel of cancer cell lines, to be used as response variables. Gene expression data, copy number data, and mutation data were provided to the participants to be used as predictor variables. A hold out set was used to score the prediction performance of each team. There were three major tasks defined:

i. Sub-challenge 1: Build a model that best predicts the gene essentiality values of thousands of genes, using the molecular and genomic characteristics/features of the cancer cell lines.

ii. Sub-challenge 2: Identify the most predictive features for each gene essentiality among a prioritized list of genes. For each prioritized gene, the aim was to select a small set of at most 10 predictive features (gene expression, copy number, and mutation), and then predict gene essentiality using only these features.

iii. Sub-challenge 3: Identify the most predictive features common for all gene essentiality values of a prioritized list of genes. For the set of all prioritized genes, the aim was to identify a single list of at most 100 shared predictive features, and then predict essentiality using only these features for all the prioritized genes.

5 Results

In the following sections, I present the results that highlight my contributions to the development of computational approaches and their application to advancement of cancer biology in general, and precision oncology in particular.

5.1 Consistency of genome-wide shRNA screens

To assess the consistency of genome-wide RNAi screens, I made use of the publicly available datasets based on genome-wide shRNA screens in large panels of cancer cell lines from different research laboratories. Project Achilles is an initiative by the Broad institute, wherein 102 cell lines from various cancer types were screened in the first phase, Achilles 2.0 (*186*), and later extended to 216 cell lines in Achilles 2.4 (*32*). COLT-Cancer (*187*) and Breast Functional Genomics (BFG) (*34*) datasets were generated by the Moffat lab and Neele lab respectively at the University of Toronto (Canada).

The Achilles projects used a genome-wide shRNA library of ~54k shRNAs, whereas the Toronto projects screened a library of ~78k shRNAs (Figure 4). However, all screens used the common library obtained from the same resource, The RNAi consortium (TRC) database. While Achilles 2.0 and COLT-Cancer measured shRNA abundances by microarray hybridization, Achilles 2.4 and BFG used NGS for the same. All the screens are similar in terms of their experimental workflow for conducting a genome-wide shRNA screen, with differences in the number of population doublings before the final shRNA abundance measurement. A substantial number of identical cell lines were also screened in between the Achilles and Toronto projects (Figure 4), making it possible to perform a quantitative assessment of the consistency between studies in terms of the shRNA-level phenotypes and gene-level dependencies in publication **I**.

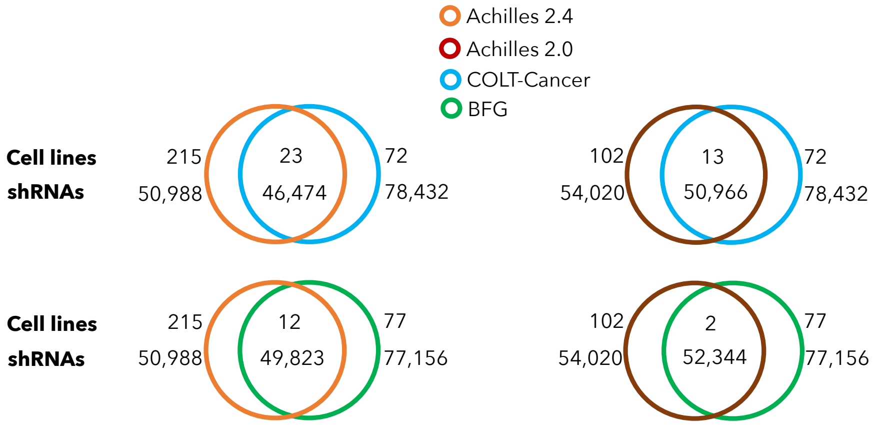


Figure 4: Overlap in shRNAs and cancer cell lines screened in the Project Achilles, COLT-Cancer and Breast Functional Genomics (BFG) screens.

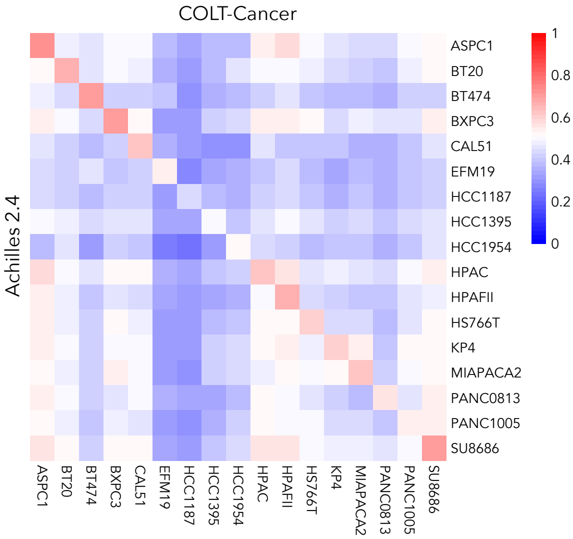


Figure 5: Heatmap of rank correlation of shRNA essentiality scores (shES) between Achilles 2.4 and COLT-Cancer projects for common set of shRNAs and cell lines.

Correlation analysis of shRNA-level phenotypes, i.e. shRNA essentiality scores (shES) for the common set of shRNAs between the identically screened cell lines in Achilles 2.4 and COLT-Cancer, revealed a moderate consistency between the two studies (mean rank correlation = 0.57) (Figure 5). Moreover, the between-study correlations between identical cell lines was systematically higher than either the intra-study or interstudy correlations between the non-identical cell lines, suggesting that the phenotypic effects of shRNAs are significantly influenced by the genetic background of the screened cell line (Figure 5). In addition, the type of platform for measurement of shRNA abundance also influenced the consistency of the screens. Average correlation between screens using microarray hybridization, Achilles 2.0 and COLT-Cancer was much lower (mean rank correlation = 0.38), than between screens using NGS, Achilles 2.4 and BFG (mean rank correlation = 0.53).

While a quantitative estimate of the shRNA-level phenotypic effects on cell proliferation is the outcome of a genome-wide shRNA screen, quantifying the gene-level dependencies of cancer cells is desired for analytical purposes and for building predictive computational models. Since shRNAs are known to exhibit off-target effects, the methods for summarization of shES scores into gene essentiality scores (geneES) can influence the accuracy of inferred genetic dependencies of cancer cells, and consequently the consistency of genome-wide shRNA screens. Methods summarizing the intended on-target activity of shRNAs, such as, RIGER, ATARiS, RSA and average gene essentiality (AGE), led to a decrease in the consistency of the Achilles 2.4 and COLT-Cancer screens, in comparison to the shES-based rank correlation estimates (Figure 6A). In contrast, the correlation of GARP-based geneES for identical cell lines did not decrease significantly. Surprisingly, the consistency between the two screens increased significantly (mean rank correlation = 0.71, $p = 8.6 \times 10^{-08}$), when analysed based on seed essentiality (seedES) scores. seedES summarizes the off-target activity of shRNAs, by averaging the shES of all shRNAs having an identical nucleotide sequence at the seed region (position 2-8) of the guide strand.

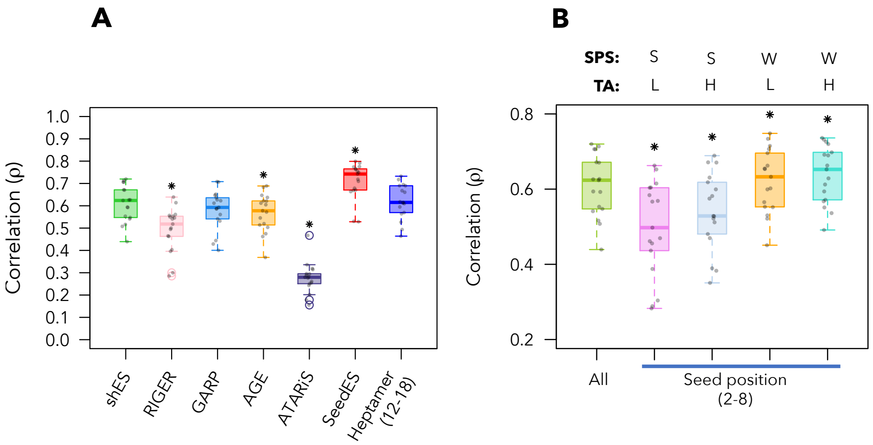


Figure 6: (A) Boxplot of rank correlations between Achilles 2.4 and COLT-Cancer screens based on shES, geneES and seedES. Asterisks indicate statistically significant differences in correlations (p < 0.05, paired t-test). geneES scores are estimated by RIGER, GARP, AGE and ATARiS methods which summarize the intended on-target effect of shRNAs. Average Gene Essentiality (AGE)-based geneES were calculated by averaging the shES scores of all shRNAs targeting an intended gene. SeedES were calculated by averaging the shES of sets of shRNAs having the same seed sequence, with set size >= 5. Hepatmer12-18ES is the average shES of shRNAs having identical sequence from positions 12-18. (B) Boxplot of rank correlations based on shES for shRNAs categorized based on their biochemical seed sequence properties: seed pairing stability (SPS) and target abundance (TA). shRNAs were categorized into combinations of strong SPS or weaker SPS, and lower TA or higher TA. Asterisks denote statistically significant differences in correlation (*p* < 0.05, paired *t*-test).

Further, it was found out that properties of seed sequences that are known to affect the off-targeting tendency of shRNAs alsoinfluenced the consistency of the screening results (Figure 6B). The sequence composition of seed region of a shRNA is known to affect its biochemical properties, such as how strongly it pairs with off-target mRNAs or many it can bind with (*86*). Seed pairing stability (SPS) is a measure of the thermodynamic stability of the seed-mRNA duplex, and target abundance (TA) is a measure of the availability of target mRNAs with a seed complementary sequence. Stronger SPS values suggest that the binding is more stable and hence having a higher likelihood of off-target effects, and vice-versa for weaker SPS values. Higher TA values mean more number of available target mRNAs and the shRNAs are titrated out, thus having milder effect on down-regulation of off-target mRNAs. The consistency of the two screens was remarkably lower for shRNAs categorized as stronger SPS and lower TA or stronger SPS and higher TA (Figure 6B).

5.2 Prediction of novel synthetic lethal interactions

Genome-wide RNAi screens can be used to identify the context-specific addictions of cancer cells, for instance, the addictions that are present in the cancer cells having a mutated driver gene while not in the wild type background. Such context-specific dependencies, also known as synthetic lethal interactions, serve as a useful principle for identifying non-direct approaches for targeted therapy. The Achilles 2.4 and COLT-Cancer studies have profiled large panels of cell lines from a wide background of lineages, which allowed us to perform statistical analyses to detect candidate genes that are robust synthetic lethal partners, i.e. differentially dependent, of frequently mutated cancer driver genes in publication **I**.

Moreover, identification of robust synthetic lethal interactions is also dependent on the accuracy of gene dependency estimates and the genetic background in which the context-specific dependency relationships exist. The accuracy of gene dependency estimates can be improved by accounting for the off-target effects of shRNAs. It was found out that removing the shRNAs with higher propensity for off-target effects, based on their seed sequence properties, from the estimation of GARP-based geneES scores led to an improved consistency between the Achilles 2.4 and COLT-Cancer screen in publication **I**. Likewise, we observed an improvement in the common number of synthetic lethal candidates identified between the two screens for several cancer driver genes (Figure 7). In addition, similar improvement in identification of genetic interaction partners of the cancer driver genes was also observed.

To test whether this approach was successful in predicting novel synthetic lethal partners, we further studied the synthetic lethal partner of PIK3CA driver oncogene that were identified only post-removal of shRNAs with higher propensity of off-target effects. Two putative synthetic lethal partners, PKN3 and HMX3, were identified as synthetic lethal hits of PIK3CA gene in both of the datasets. Knockout of these genes using CRISPR/Cas9 in isogenic MCF10A cell lines having two different PIK3CA driver mutations, E545K and H1074R, led to a systematic decrease in proliferation of the cells, hence confirming the robust synthetic lethal nature of these genes with PIK3CA.

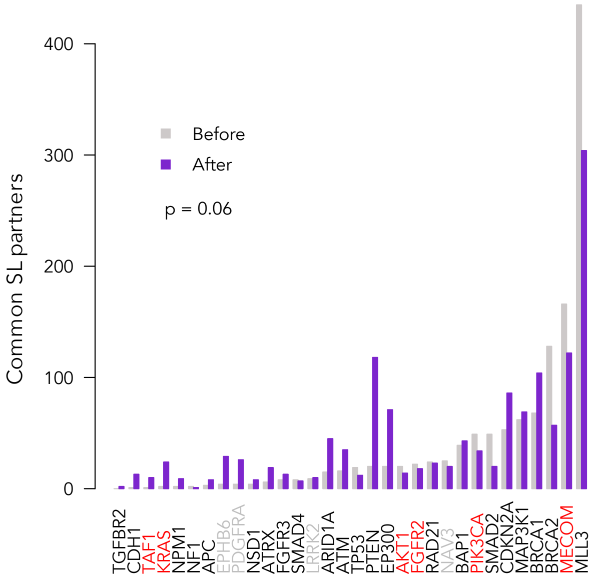


Figure 7: Systematic increase in overlap of synthetic lethal candidate partners of several cancer driver genes after removing the shRNAs with higher propensity of off-target effects. P-value is calculated based on a Wilcoxon signed rank test.

5.3 Predicting gene dependencies in cancer cell lines

Predicting gene essentiality profiles of cancer cells using genetic and molecular profiles can provide biological insights into the systems-level genetic interactions and dependencies across cancer cells. The task in the Broad-DREAM essentiality prediction challenge (publication **II**) was to build machine learning models that can best predict the gene dependency profiles of cancer cells using their genomic and molecular profiles.

For sub-challenge 2 and 3, in which the task was to predict gene essentialities of selected 2467 genes, the MT-GRLS model was implemented (Figure 8). MT-GRLS performs step-wise greedy forward selection by adding at each step the feature whose addition leads to the largest increase in leave-one-out cross-validation performance over all the target genes. The algorithm is highly scalable having a linear time training complexity, and it directly optimizes the predictive performance of the learned model subject to the budget constraints, making it ideal for these sub-challenges.

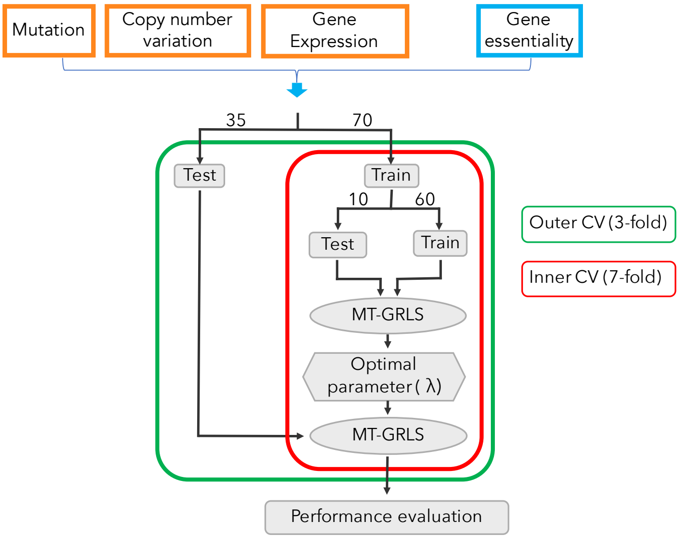


Figure 8: Schema for learning the MT-GRLS model for predicting gene dependencies of cancer cell lines. Training data for 105 cell lines were used in a nested cross-validation setting to select the best model parameters (λ) with maximum predictive performance in the training data. The final selected parameter was used in the test data of 44 cell lines.

Before applying the MT-GRLS, we implemented data-preprocessing steps to reduce the number of redundant and non-informative genetic and molecular features. For the genes with identical CNV profile across all the cell lines, we used the first non-duplicate row to reduce the number of duplicate features (i.e. identical CNV profiles). Missing mutation status data were treated as wild type and the genes with zero-variance in their mutation profiles across the cell lines were removed. Finally, all the features (i.e. gene expression, CNV and mutation data matrices), as well as the gene essentiality profiles (response variables) were normalized to zero-mean and unit variance. We evaluated the performance of MT-GRLS model internally, using a nested-cross validation (CV) approach (Figure 8). First, we applied the model over a range of regularization parameters with 7-fold inner CV to select the most predictive regularization parameter. Then, the predictive accuracy of the model was evaluated using a 3-fold outer CV loop. The nested CV provided an accurate estimate of the prediction accuracy on the independent test set.

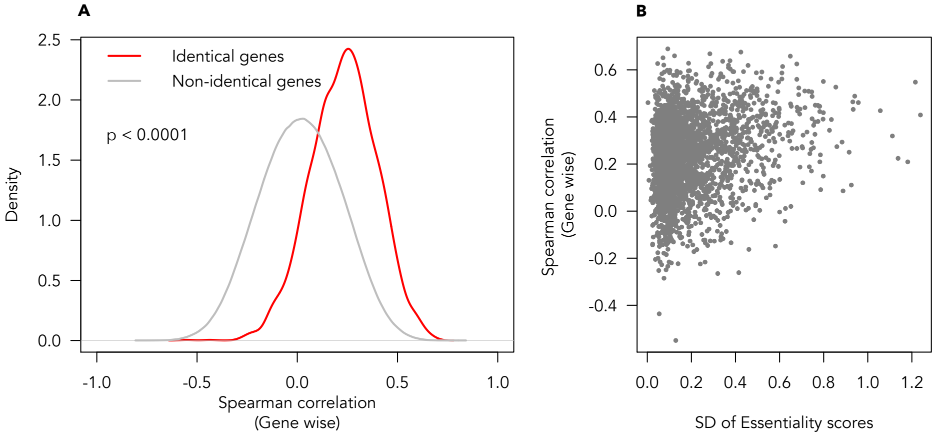


Figure 9: Rank correlation between predicted and observed gene essentiality scores in sub-challenge 3. (A) Density distributions of the rank correlations. Red line indicates the correlation between the identical genes, and the gray line indicates the correlation between the non-identical genes, which is used as a baseline prediction performance measure. P values are obtained from a Wilcoxon rank-sum test. (B) Scatterplot of rank correlation between predicted gene essentialities and their standard deviations. Genes with higher standard deviations had higher prediction accuracy.

The final prediction model was based on the best regularization parameter learned from the complete training dataset using 7-fold CV. It was found out that the prediction performance of the MT-GRLS did not benefit from any prior filtering of the features, perhaps due to its efficient feature selection procedure. We evaluated the predictive performance of the trained models by rank correlation between the predicted and observed gene essentiality profiles. Our MT-GRLS model was the top-peforming method in the sub-challenge 3, where its average correlation for predicted gene essentialities was 0.23, which was significantly higher than the baseline correlation observed between the non-identical genes, suggesting that genetic datasets have substantial predictive power (Figure 9A). Further, we observed that individual genes whose essentiality scores were more variable across the cell lines were the ones for which predictive accuracy was also high, suggesting that each genes feature contributes substantially to the overall predictive performance of the models (Figure 9B). Gene set enrichment analysis revealed that the highly predictable genes were enriched for basic cellular processes and functions (Table 2), such as proteasome, spliceosome, DNA damage and repair, cell cycle, oxidative phosphorylation and targets of the transcription factors E2F and MYC.

Gene Set	**FDR q-value**
Hallmark Myc targets V1	1.79×10^{-20}
KEGG Spliceosome	2.00×10^{-17}
KEGG Proteasome	2.98×10^{-16}
Hallmark E2F targets	5.44×10^{-15}
Hallmark G2M checkpoint	6.13×10^{-14}
Hallmark DNA repair	8.82×10^{-13}
KEGG Cell cycle	2.16×10^{-10}
Hallmark Oxidative phosphorylation	6.71×10^{-10}
KEGG Ubiquitin mediated proteolysis	5.58×10^{-09}
Hallmark Mitotic spindle	5.58×10^{-09}
KEGG Ribosome	5.58×10^{-09}
Hallmark MTORC1 signalling	4.22×10^{-07}
KEGG Focal adhesion	4.22×10^{-07}
KEGG Oocyte meiosis	9.35×10^{-07}
Hallmark PI3K-ATK-MTOR signalling	4.69×10^{-06}

Table 2: Gene set enrichment analysis of the highly predictable genes (rank correlation ≥ 0.4) with MSigDB gene sets from KEGG and Hallmark collection.

Moreover, we observed that out of the top 100 features selected for sub-challenge 3, there were no mutation features and only two features from the CNV data, suggesting that gene expression has higher information content for gene essentiality prediction, and that CNV and mutation profiles may provide partly redundant information. Analysis of the top 100 features for the sub-challenge 3 provided by the other teams revealed that EIF2C2 gene was frequently selected by several prediction models. Since EIF2C2 gene is a part of the RNAi machinery, this suggests that the expression levels of the components of RNAi machinery can influence the phenotypic outcomes of a RNAi screen. Gene set enrichment analysis of the top 100 selected features revealed also enrichment for genes involved in epithelial-mesenchymal transition, suggesting that cell state phenotypic information is predictive of gene essentialities.

5.4 Predicting drug response of cancer stem cells using gene signatures

Cancer stem cells are known to exhibit exquisite sensitivity to the small molecule inhibitor, salinomycin, however there is a lack of mechanistic understanding of its precise mode of action. We reasoned that molecular insights gained from experimental studies can be used to derive a gene expression signature to predict the response of cancer cells to salinomycin, and further aid in identifying groups of patients or tumor types that would benefit the most from salinomycin therapy.

We observed that salinomycin treatment of cells was associated with disruption of the KRAS nanoscale membrane organization by altering the distribution of phosphatidyl serine (PS), eventually leading to decreased signalling output from KRAS nanoclusters due to reduced effector recruitment. Moreover, overexpression of caveolin decreased the sensitivity of cells to salinomycin by affecting the membrane organization (see publication **III** for details). This suggests that gene expression state of known modulators of KRAS nanoscale membrane organization can influence the drug response. To gain further insights into the gene expression signature associated with KRAS nanoscale membrane organization, we utilized the ESTOOLS database (*180*) to find genes that are correlated with its known modulators. Based on the 13 genes that were identified, the gene expression signature classified embryonic stem cells separately from the fibroblasts, suggesting that the signature was also associated with stemness property (Figure 10).

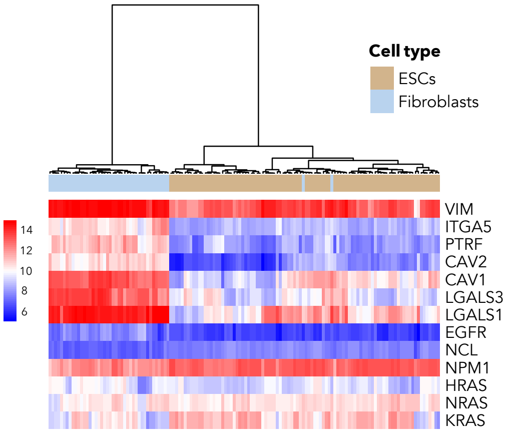


Figure 10: Clustering of embryonic stem cells (ESCs) and fibroblasts based on the gene expression of known modulators and correlated genes (VIM, ITGA5 and CAV2). Gene expression data of Metaset 1 from ESTOOLS database is presented as heatmap.

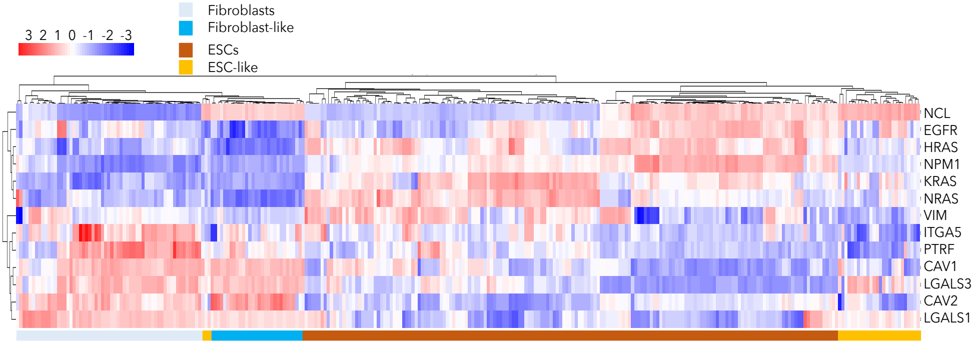


Figure 11: Unsupervised clustering of selected cancer cell lines identified as ESC-like and fibroblast-like based on correlation with the KRAS nanoclustering associated gene expression signature. Gene expression data from CCLE and ESTOOLs were quantile-normalized and scaled.

To study the stemness property associated with KRAS nanoclustering in cancer cells, we further identified cancer cell lines that were correlated with gene expression signature of the 13 genes in ESCs and fibroblasts. As expected, we observed that the ESC-like cell lines clustered with the ESCs and the fibroblast-like cell lines clustered with the fibroblasts (Figure 11). Drug sensitivity profiling revealed that the ESC-like cell lines were more sensitive to salinomycin whereas less responsive to staurosporine compared to fibroblast-like cell lines, suggesting that the gene-expression signature is capable of predicting the stemness property of cancer cell lines and also its response to a CSC inhibitor (Figure 12).

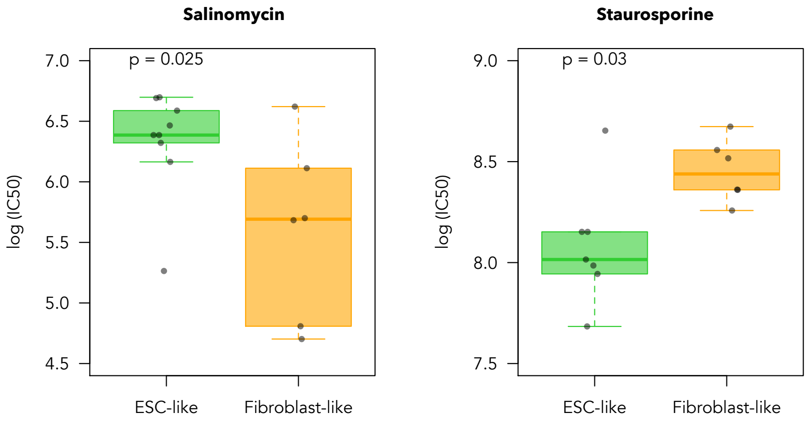


Figure 12: Drug response levels of ESC-like and fibroblast-like cancer cell lines to salinomycin and staurosporine. Logarithm of IC50 values were obtained from a drug dose response curve. p values were obtained by one-side Wilcoxon rank sum test.

We further hypothesized that patient tumor samples exhibiting the gene expression signature associated with stemness property should present differences in their clinical characteristics. To assess that, we performed correlation analysis to identify the patient tumor samples in The Cancer Genome Atlas (TCGA) dataset that were displaying ESC-like and fibroblast-like gene expression signature. Interestingly, we found that ESC-like patient samples were associated with lower survival probability, when compared to non-ESC like samples. As expected, fibroblast-like samples did not show the same difference (Figure 13), suggesting that patient tumors that are more cancer stem cell like are more aggressive.

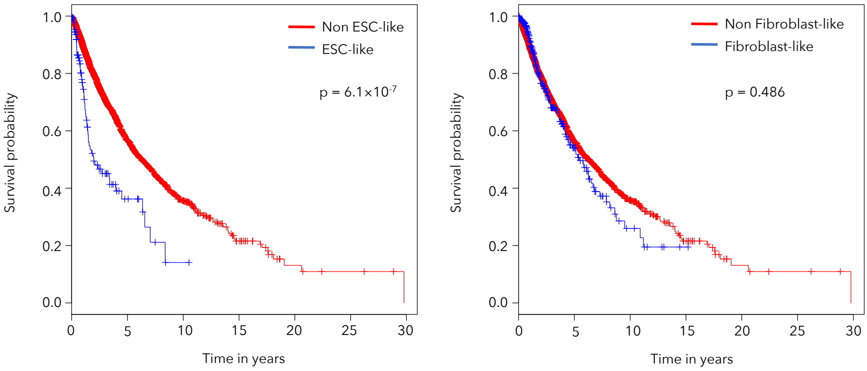


Figure 13: Survival analysis of patient tumor samples from TCGA, defined as ESC-like (rank correlation >= 0.6) and Non ESC-like (rank correlation <= 0.2) and Fibroblast-like like (rank correlation >= 0.6) and Non fibroblast-like (rank correlation <= 0.2) based on correlation with gene expression signature with ESCs and fibroblasts.

6 Discussion

The rapid advancements in high-throughput techniques have now made it possible to molecularly characterize large number of patient tumors, and large-scale genomic and functional profiles are routinely being generated. Such datasets hold immense potential to reveal novel genes driving cancer, biomarkers with prognostic value, and also identify promising targets for drug treatment. But the 'big data' nature of these highly complex datasets require concurrent development of computational models and data analysis strategies to be able to mine useful knowledge and unlock the potential of the information content that is latent in such datasets. This thesis presents computational and analytical approaches to extract potentially useful information by integrating genomic and functional profiles of cancer cells.

Publication **I** demonstrates how in-depth information on the mechanistic properties of shRNAs can be utilized to remove noise from genome-wide shRNAs screen datasets in post-screening analysis scenario. The study particularly aimed to explore means to increase the consistency between genome-wide shRNA screens, so that these lessons can be incorporated in the designing of future genome wide shRNA screens. Reassuringly, the study found moderate consistency between the genome-wide shRNA screens, suggesting that although there is a considerable amount of noise in the data, it still has the potential to yield promising results. The study demonstrated that consistency between shRNA screens is significantly higher for the seed mediated off target effects. As observed in a previous study [29], we also find that the consistency between datasets increases significantly based on seed essentiality scores.

While it is expected that the specific phenotypic effects of each shRNA within a shRNA family might differ in terms of the target profile of down-regulated off-target genes, averaging overall the constituent shRNAs members in a family was found to be indicative of the phenotypic effects of the shared off-target profile of genes. This could explain the observed increase in consistency between the screens. From the observations based on our study, we propose that saturating the seed sequence space by sampling over multiple shRNAs having the same seed sequence while designing genome wide shRNA libraries is a good approach to accurately estimate seed level essentiality scores. This in turn can be used to model the off-target genes based on seed sequence complementarity which may allow us to derive more accurate gene essentiality scores. Computational methods modelling the seed-mediated effects that have been implemented previously to discern the off-target genes in RNAi screens (*188-191*), however their shortcoming is that they are unable to provide gene essentiality scores for all genes screened. By focussing on methods that can be implemented easily for derivation of gene essentiality estimates, this study adopted a simplistic approach by enriching the shRNAs with on-target activity.

From a practical point of view, Publication **I** provides a straightforward approach that can be incorporated in the analysis of existing genome wide RNAi screening datasets to extract the most accurate biological information out of them. The study identified 'bad quality' shRNAs with higher propensity of off-target effects based on determinants of targeting proficiency of miRNAs, i.e. SPS and TA. Reporter activity studies have previously shown that a strong pairing leads to stronger repression of bound target and hence proficient down-regulation of off-target transcripts [25]. SPS is a measure of the thermodynamic stability [24], a proxy for standard free energy change (Δ*G*) for the formation of the seed duplex. Predicted SPS has been calculated after taking into account several biochemical parameters and base composition [27]. More negative values of free energy change, i.e. stronger SPS, suggests that seed duplex is more stable, whereas higher values, i.e. weaker SPS, suggest less stable pairing. Further, this study demonstrated the quantitative effect of these bad quality shRNAs on the loss of consistency of genome-wide shRNA screens. We were able to show that removing the bad quality shRNAs from post-processing led to better estimates of gene dependency scores using conventional methods for summarizing shRNA level scores to gene level essentiality scores. In the future, computational models incorporating the biochemical properties of seed sequences should be developed to derive more accurate estimates of gene essentiality.

We also demonstrated that performing such post-processing can help in identifying novel synthetic lethal partners of cancer driver genes, which we also validated using a complementary CRISPR/Cas9 knockout screen. One of the important areas of applications of genome-wide RNAi screens is to identify dependencies of cancer cells in a certain genetic background that can provide interesting targets for anticancer treatment. In publication **I**, we showed how one can extract information on robust synthetic lethal interactions partners from noisy genome-wide shRNA screens. Moreover, analysing multiple datasets on a large panel of cell lines from diverse lineages and cell types is a useful way to account for the genetic heterogeneity known to exist in tumors and identify 'pan-cancer' synthetic lethal interactions.

While our approach to identifying synthetic lethal partners is based on the conventional viewpoint of differential dependencies in the mutated and wild type cell lines, other paradigms for defining synthetic lethal interactions also exist. For instance, synthetic dosage lethality is a type of genetic interaction in which the upregulation in mRNA or protein levels of one partner gene and the loss-of-function of the other partner gene results in a lethal phenotype (*161*). Synthetic lethal interactions are also known to be condition-specific, such as being dependent on the cellular state, metabolic state, genetic background or tumor microenvironment (*161*). Hence, synthetic lethal interactions observed under laboratory conditions in cancer cell lines may not be relevant in the context of overall human physiology, and thus clinical responses may not be observed.

The CRISPR/Cas9 system has recently emerged as an alternative to RNAi technology for high-throughput loss-of-function genetic screening. Similar to genome-wide RNAi libraries, several genome-wide CRISPR/Cas9 single guide RNA (sgRNA) libraries are nowadays available for functional genetic screening (*192-195*). A better understanding of the relative strengths and limitations of the two technologies would be of prominent interest to the biomedical research community. Evers et al. (*196*) and Morgens et al. (*197*) recently conducted a systematic comparison by targeting a reference set of known essential and non-essential genes to assess the relative efficiency of the two approaches; however, the two studies differ in their conclusions. The current perspective is shaping up in favor of CRISPR-based screens, as these are expected to produce more robust and sensitive phenotypes; this view was also supported by the two comparative studies, although the Evers study (*196*) was more positive about the superiority of the CRISPR technology, whereas the Morgens study (*197*) concluded that both technologies have their respective strengths and limitations. Understanding the factors affecting sgRNA activity will be crucial in assessing the relative performance of CRISPR and RNAi screens, with the aim at defining the best practices for loss-of-function screening and designing the most efficient genome-wide sgRNA and shRNA libraries. Off-target effects have also been shown in CRISPR/Cas9 screens (*198*), and several extrinsic factors, such as the expression of Cas9 (*199*), sgRNA sequence properties (*200*), targeted region of protein domains, DNA accessibility and local architecture of the genomic region of the target locus, may also affect the performance of CRISPR screens.

Publication **II** demonstrated how genomic features of cancer cell lines can be used to predict their functional gene essentiality profiles by using machine learning models. With the availability of high-throughput technologies, it has become easier to profile larger number of tumors and generate copious amounts of data representing their molecular characteristics. To make sense of these datasets, computational models are needed to integrate the multiple layers of information for identifying novel ways of treating cancer. The Broad-DREAM gene essentiality prediction challenge demonstrated a novel approach in which a community effort is leveraged for solving important biomedical questions, by establishing benchmark models for prediction tasks. We developed MT-GRLS model in sub-challenge 3, demonstrating that the best performing method selects sparse panel of genomic features that are predictive of gene essentialities of multiple genes. MT-GRLS exploits multitask learning, which leverages information that is shared across multiple variables, and therefore increases the statistical power of the inference problem.

A consistent finding in publication **II** was that gene expression data contain more predictive information compared to other molecular datasets, as has been observed also in other DREAM challenges (*169, 201-203*). Gene expression features were also the most prominently selected top 100 features in sub-challenge 3. This may reflect the fact that most of the predictive models are well suited to incorporate continuous variables, whereas extracting predictive information from categorical datatypes, such as mutations and copy number variations, has proved more challenging for the current models. Analysis of the frequently selected gene expression features revealed that expression levels of EIF2C2 has significant predictive power of the gene essentiality scores. This suggests that the functional state of the RNAi machinery influences the efficiency of knockdown and thus the inferred dependency scores. Future predictive models should take this into account, and moreover consider that genome-wide RNAi screens based phenotypes need to be interpreted cautiously. More importantly, this information should be used in post-processing of genome-wide RNAi screens to estimate accurate gene dependency scores. Moreover, the most predictive gene expression signatures were enriched for genes involved in epithelial-mesenchymal transition genes indicating that the phenotypic cell state are highly informative of the gene essentialities. Perhaps this reflects the previous observation that cell lines cluster into two major groups based on gene expression data that correspond to the epithelial and mesenchymal states.

The sub-challenge 3 prediction task was restricted to the use of genomic and molecular information only, namely mutation, CNV and gene expression, which might explain the modest average performance of the prediction models. Combining information from multiple other datatypes, such as epigenome, proteome, metabolome and other molecular portraits of cancer cell lines, could potentially contribute to enhanced prediction performance. Also, addition of prior biological knowledge such as biological pathways and processes can improve the prediction performance, as has been observed previously (*169, 174*). Moreover, systems biology based integrative models that take into account the different types of molecular information, and the network and signalling properties of genes, can further bring in additional information that are predictive of gene essentialities.

Publication **III** explored the link between stemness property and nanoscale membrane organization of KRAS. Cancer stems cells have been linked to EMT transition, and it is likely that KRAS signalling also contributes to EMT via Wnt pathway. Additionally, the study demonstrates how the mechanistic understanding of affectors of KRAS nanoclustering can be coupled with computational analysis of gene expression data to build an expression signature predictive of response to CSC inhibitors. Importantly, the gene expression signature can also be applied in stratifying patients that are more likely to respond to salinomycin or other CSC inhibitors. The enrichment of breast cancer subtypes in the tumor-types that were ESC-like is in agreement with previous studies which identified salinomycin as a CSC inhibitor (*142*). Acute myeloid leukemia cancer-type was also enriched in the ESC-like group, corroborating previous results indicating link between stem cell expression signature and survival outcomes (*204*).

In conclusion, this thesis demonstrates that computational approaches to integrate functional and genomic datasets of cancer cell lines can be useful in understanding cancer biology and guide further translational efforts. Prudent implementation of relevant biological information to the analysis of genome-wide RNAi screen datasets can be useful in reducing the noise inherent in these datasets. Ultimately this leads to more accurate dependency maps of cancer cells, and therefore may reveal potential therapeutic targets for cancer treatment. The study also demonstrates that predictive models can be built for gene dependency profiling of cancer cell lines. Predictive modelling basd on integrated genomic and functional datasets can yield insightful knowledge on the molecular characteristics of cancer cells, such as the predictive value of EMT phenotype and the biological processes whose dependencies can be predicted more accurately. Additionally, the study indicates that the transcriptomic landscape has high predictive power for the functional landscape of cancer cells. The thesis also demonstrates the power of coupling computational approaches with biological hypotheses in predicting drug response phenotypes and identifying clinically relevant information about patient tumors.

As a future development, genome-wide loss-of-function screens based on complementary CRISPR/Cas9 knockouts will be likely useful in estimating more accurate genetic dependencies of cancer cell lines. Computational methods to reduce noise in loss-of-function screens, similar to those developed in this thesis, should lead to further improvements in accuracy of predictive models of gene dependency scores based on genomic datasets. Also, incorporating information of proteomic and epigenomic landscapes of cancer cell lines could lead to improvement in the predictive accuracy of genetic dependencies. Loss-of-function and molecular profiling in more advanced cancer cell line models, such as those based on 3D organoids that recapitulate the tumour features more realistically, may further provide better ways to find novel targets. In addition, there is a need to develop computational methods that are able to quantitatively account for the mechanistic details and several levels of biological organization; such as the signalling and pathway level interactions and network-level properties of genes and proteins. A holistic systems-biology based modelling approach may lead to a better understanding of the biology of cancer and will be useful in identifying promising targets for cancer treatment.

7 Acknowledgements

Sincere thanks to Prof. Tero Aittokallio for being an amazing supervisor. I have had a fantastic opportunity to learn so much from you. Thank you so much for giving me the support and independence to work on my own projects as well. Considering my background has been more biology oriented, it was a big change in field to come into your lab and initial days were quite hard. But, you have been super supportive and always inspired me to take more courses. Your ease of availability for discussion on project work and your easy-going nature has been very helpful, especially for someone like me who has a long way to go to get organized. Your efficiency in writing and the feedbacks that you provide have been a thing of inspiration for me. Thank you also for the lunch time discussions that we use to have, which varied in topics and were not always “sciency”. Without your support and mentorship, I would not have managed to get this thesis work done in time.

Sincere thanks to my co-supervisor Dr. Jing Tang. Your mentorship has been very crucial for me to get a better perspective on the “computational and quantitative” way of thinking. Discussions with you have been extremely helpful for my projects and I learned a lot from your approach of doing research work. A big thanks to Dr. Krister Wennerberg for your fantastic mentorship, ease of availability for discussions and important inputs into my projects. You have been a figure of inspiration for me and I have gained confidence on my own scientific ideas based on discussions with you. Sincere gratitude to my co-authors: Dr. Sergey Kuznetsov, Prof. Daniel Abankwa, Dr. Arafath Kaja Najumudeen, Prof. Samuel Kaski, Dr. Kerstin Bunte, Dr. Tapio Pahikkala, Dr. Antti Airola, Prof. Olli Kallioniemi, Dr. Sara Kangaspeska and Susanne Hultsch who have been very wonderful to work with. Without your collaborative contributions, I would not be able to accomplish and produce this body of work.

This project has been carried out over a timeline of 5 years at the Institute for Molecular Medicine Finland (FIMM). I would like to thank all my colleagues at FIMM to make it a great place to work, with a vibrant community of students and scientists and extremely good facilities and above all a hierarchy-less environment. I would like to thank Dr. Gretchen Repasky, our Research Training Coordinator at FIMM and the coordinator of the International FIMM-EMBL PhD programme. An opportunity to undergo rotations in several labs at

FIMM before committing to this PhD work was very important for me to get a perspective on what questions interest me. Besides, Gretchen has been super supportive in helping the student community at FIMM to be active. Being part of organizing Think Different Seminar Series, Nordic EMBL conferences, PhD symposium, FIMM PhD and Postdoc council visits allowed me to explore other sides of scientific work life, which has been an extremely useful experience.

Heartfelt thanks to the current and previous Group Aittokallio lab members: Dr. Petteri Hintsanen, Dr. Bhagwan Yadav, Dr. Abhishek Gupta, Dr. Suleiman Ali Khan, Dr. Laxmana Yetukuri, Dr. Ammad-ud-din Mohammed, Dr. Lu Cheng, Dr. Anil Giri, Dr. Zia Ur Rehman, Yevhen Akimov, Sanna Timonen, Balaguru Ravikumar, Himanshu Chheda, Anna Cichonska, Agnieszka Szwajda, Liye He, Alexander Ianevski and Zaid Alam. You all have been great co-workers and I have enjoyed our group retreats, lab meeting discussions and lab lunches with you. It has been a pleasure to work with you all and you have contributed to my growth as a researcher. I would especially like to thank Dr. Gopal Peddinti for his brilliant mentorship and unwavering persistence and support. Him being in the same cubicle was crucial for me to learn and get a quick word on a script gone wrong, or a quick chat on never-ending questions on computational modeling, algorithms and statistics. Your mentoring, in addition to the discussions we had on science and philosophy have been extremely useful in broadening my perspective.

I would like to thank my Graduate school, Integrative Life Sciences (ILS) for providing me with the funding for my studies, and also a lot of opportunities for networking with fellow PhD students and participate in variety of courses. I would especially like to thank ILS coordinator, Dr. Erkki Raulo for being so supportive of students, organizing the epic parties and also for setting up the ILS Student Council. Wholehearted thanks to the dynamic, upbeat and vivacious friends from the council: Behnam, Elina, Geri, Heidi, Heini, Kornelia, Maarja, Mridul, Sigurdur and Tuomo. I want to thank the members of The Science Basement: Lea, Barun, Chiara, Dmitrios, Petra, Susanne and Ekaterina. It has been a pleasure working with you guys and building this student organization with a zeal for science communication. I would also like to thank my dearest Nepalese friends: Shishir, Kul, Anil, Preeti, Sawan, Arya, Om, Rubina, Deepak, Shruti, Abhishekh, Bhagwan, Ashwini, Disha, Himanshu, Balaguru, Swapnil and Christian who have been an important part of my social experience in Helsinki. A special thanks to Prson

Gautam with whom I have worked, travelled, been gym partners and taken part in numerous social gatherings. Also, Barun Pradhan for being such a great flatmate and companion. I also want to thank my friends: Anubha and Gopal who were an immense source of inspiration before coming to Helsinki.

Lastly, I would like to thank my family members for their unflinching support, faith and love for me. Without whom I would not have had the confidence to learn and do anything.

8 References

1. WHO fact sheet http://www.who.int/mediacentre/factsheets/fs297/en/.
2. T. Ngoma, World Health Organization cancer priorities in developing countries. *Annals of Oncology* **17**, viii9-viii14 (2006).
3. World Cancer Report 2014.
4. The Biology of Cancer. Weinberg RA. 2nd ed. New York: Garland Science; 2014.
5. D. Stehelin, H. E. Varmus, J. M. Bishop, P. K. Vogt, DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170-173 (1976).
6. D. Hanahan, R. A. Weinberg, The Hallmarks of Cancer. *Cell* **100**, 57-70 (2000).
7. D. Hanahan, Robert A. Weinberg, Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646-674 (2011).
8. Levi A. Garraway, Eric S. Lander, Lessons from the Cancer Genome. *Cell* **153**, 17-37 (2013).
9. J. G. Paez *et al.*, EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. *Science* **304**, 1497 (2004).
10. T. J. Lynch *et al.*, Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. *New England Journal of Medicine* **350**, 2129-2139 (2004).
11. Y. Samuels *et al.*, High Frequency of Mutations of the PIK3CA Gene in Human Cancers. *Science* **304**, 554 (2004).
12. H. Davies *et al.*, Mutations of the BRAF gene in human cancer. *Nature* **417**, 949-954 (2002).
13. C. Sawyers, Targeted cancer therapy. *Nature* **432**, 294-297 (2004).
14. S. V. Sharma, J. Settleman, Oncogene addiction: setting the stage for molecularly targeted cancer therapy. *Genes & Development* **21**, 3214-3231 (2007).
15. F. S. Collins, A. D. Barker, Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am* **296**, 50-57 (2007).
16. International network of cancer genome projects. *Nature* **464**, 993-998 (2010).
17. D. Dickson, Wellcome funds cancer database. *Nature* **401**, 729-729 (1999).
18. B. Vogelstein *et al.*, Cancer Genome Landscapes. *Science* **339**, 1546 (2013).
19. W. C. Hahn *et al.*, Integrative genomic approaches to understanding cancer. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1790**, 478-484 (2009).
20. S. E. Moody, J. S. Boehm, D. A. Barbie, W. C. Hahn, Functional genomics and cancer drug target discovery. *Curr Opin Mol Ther.* **12**, 284-293 (2010).

21. T. P. Howard *et al.*, Functional Genomic Characterization of Cancer Genomes. *Cold Spring Harbor Symposia on Quantitative Biology* **81**, 237-246 (2016).
22. J. Mullenders, R. Bernards, Loss-of-function genetic screens as a tool to improve the diagnosis and treatment of cancer. *Oncogene* **28**, 4409-4420 (2009).
23. O. Shalem, N. E. Sanjana, F. Zhang, High-throughput functional genomics using CRISPR–Cas9. **16**, 299 (2015).
24. D. E. Root, N. Hacohen, W. C. Hahn, E. S. Lander, D. M. Sabatini, Genome-scale loss-of-function screening with a lentiviral RNAi library. *Nat Meth* **3**, 715-719 (2006).
25. S. V. Sharma, D. A. Haber, J. Settleman, Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer.* **10**, 241 (2010).
26. T. Pemovska *et al.*, Individualized Systems Medicine (ISM) strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discovery*, (2013).
27. J. Barretina *et al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-307 (2012).
28. F. Iorio *et al.*, A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754 (2016).
29. M. J. Garnett *et al.*, Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575 (2012).
30. B. Seashore-Ludlow *et al.*, Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery* **5**, 1210-1223 (2015).
31. M. G. Rees *et al.*, Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* **12**, 109-116 (2016).
32. G. S. Cowley *et al.*, Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data* **1**, 140035 (2014).
33. E. R. McDonald, III *et al.*, Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**, 577-592.e510 (2017).
34. R. Marcotte *et al.*, Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance. *Cell* **164**, 293-309 (2016).
35. J. Campbell *et al.*, Large-Scale Profiling of Kinase Dependencies in Cancer Cell Lines. *Cell Reports* **14**, 2490-2501 (2016).
36. A. Tsherniak *et al.*, Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e516 (2017).
37. J. L. Y. Koh *et al.*, COLT-Cancer: functional genetic screening resource for essential genes in human cancer cell lines. *Nucleic Acids Research* **40**, D957-D963 (2012).
38. C. J. Echeverri *et al.*, Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat Meth* **3**, 777-779 (2006).

39. W. G. Kaelin, Use and Abuse of RNAi to Study Mammalian Gene Function. *Science* **337**, 421 (2012).
40. B. Haibe-Kains *et al.*, Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389-393 (2013).
41. G. J. Hannon, RNA interference. *Nature* **418**, 244-251 (2002).
42. T. M. Rana, Illuminating the silence: understanding the structure and function of small RNAs. *Nat Rev Mol Cell Biol.* **8**, 23 (2007).
43. J. C. Carrington, V. Ambros, Role of MicroRNAs in Plant and Animal Development. *Science* **301**, 336 (2003).
44. Y. Dorsett, T. Tuschl, siRNAs: applications in functional genomics and potential as therapeutics. *Nat Rev Drug Discov.* **3**, 318 (2004).
45. S. E. Mohr, J. A. Smith, C. E. Shamu, R. A. Neumüller, N. Perrimon, RNAi screening comes of age: improved techniques and complementary approaches. *Nat Rev Mol Cell Biol* **15**, 591 (2014).
46. K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, E. V. Koonin, A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**, 7 (2006).
47. A. Bolotin, B. Quinquis, A. Sorokin, S. D. Ehrlich, Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561 (2005).
48. P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-170 (2010).
49. O. Shalem, N. E. Sanjana, F. Zhang, High-throughput functional genomics using CRISPR-Cas9. *Nat Rev Genet* **16**, 299-311 (2015).
50. M. Boutros, J. Ahringer, The art and design of genetic screens: RNA interference. *Nat. Rev. Genet.* **9**, 554 (2008).
51. C. J. Echeverri, N. Perrimon, High-throughput RNAi screening in cultured cells: a users guide. *Nat. Rev. Genet.* **7**, 373 (2006).
52. R. Bernards, T. R. Brummelkamp, R. L. Beijersbergen, shRNA libraries and their use in cancer genetics. *Nat Meth* **3**, 701-706 (2006).
53. T. R. Brummelkamp, R. Bernards, New tools for functional mammalian cancer genetics. *Nat. Rev. Cancer* **3**, 781 (2003).
54. D. Sims *et al.*, High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biology* **12**, R104 (2011).
55. T. R. Brummelkamp *et al.*, An shRNA barcode screen provides insight into cancer cell vulnerability to MDM2 inhibitors. *Nat Chem Biol* **2**, 202 (2006).
56. K. Berns *et al.*, A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431-437 (2004).
57. V. N. Ngo *et al.*, A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* **441**, 106-110 (2006).
58. F. D. Sigoillot, R. W. King, Vigilance and Validation: Keys to Success in RNAi Screening. *ACS Chemical Biology* **6**, 47-60 (2011).
59. A. L. Jackson *et al.*, Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.* **21**, 635 (2003).

60. X. Lin *et al.*, siRNA-mediated off-target gene silencing triggered by a 7 nt complementation. *Nucleic Acids Research* **33**, 4527-4535 (2005).
61. B. Haley, P. D. Zamore, Kinetic analysis of the RNAi enzyme complex. **11**, 599 (2004).
62. A. Birmingham *et al.*, 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat Meth* **3**, 199-204 (2006).
63. A. L. Jackson *et al.*, Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA* **12**, 1179-1187 (2006).
64. E. M. Anderson *et al.*, Experimental validation of the importance of seed complement frequency to siRNA specificity. *RNA* **14**, 853-861 (2008).
65. D. P. Bartel, MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**, 215-233 (2009).
66. S. E. V. Linsen, B. B. J. Tops, E. Cuppen, miRNAs: small changes, widespread effects. *Cell Res* **18**, 1157-1159 (2008).
67. V. Agarwal, G. W. Bell, J.-W. Nam, D. P. Bartel, Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
68. A. Franceschini *et al.*, Specific inhibition of diverse pathogens in human cells by synthetic microRNA-like oligonucleotides inferred from RNAi screens. *Proc Natl Acad Sci U S A.* **111**, (2014).
69. X. Lin *et al.*, `Seed' analysis of off-target siRNAs reveals an essential role of Mcl-1 in resistance to the small-molecule Bcl-2//Bcl-XL inhibitor ABT-737. *Oncogene* **26**, 3972-3979 (2006).
70. F. D. Sigoillot *et al.*, A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. *Nat Meth* **9**, 363-366 (2012).
71. F. D. Bushman *et al.*, Host Cell Factors in HIV Replication: Meta-Analysis of Genome-Wide Studies. *PLOS Pathogens* **5**, e1000437 (2009).
72. C. Scholl *et al.*, Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell* **137**, 821-834 (2009).
73. C. Babij *et al.*, STK33 Kinase Activity Is Nonessential in KRAS-Dependent Cancer Cells. *Cancer Research* **71**, 5818 (2011).
74. T. Luo *et al.*, STK33 kinase inhibitor BRD-8899 has no effect on KRAS-dependent cancer cell viability. *Proceedings of the National Academy of Sciences* **109**, 2860-2865 (2012).
75. B. Bhavneet, D. Hakim, Systematic Analysis of RNAi Reports Identifies Dismal Commonality at Gene-Level and Reveals an Unprecedented Enrichment in Pooled shRNA Screens. *Combinatorial chemistry & high throughput screening* **16**, 665-681 (2013).
76. D. H. Bhinder B, A Decade of RNAi Screening: Too Much Hay and Very Few Needles. *Drug Disc World* **14**, 31-41 (2013).
77. A. Birmingham *et al.*, Statistical methods for analysis of high-throughput RNA interference screens. *Nat Meth* **6**, 569-575 (2009).

78. R. Konig *et al.*, A probability-based approach for the analysis of large-scale RNAi screens. *Nat Meth* **4**, 847-849 (2007).
79. S. F. Yan, H. Asatryan, J. Li, Y. Zhou, Novel Statistical Approach for Primary High-Throughput Screening Hit Selection. *Journal of Chemical Information and Modeling* **45**, 1784-1790 (2005).
80. B. Luo *et al.*, Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences* **105**, 20380-20385 (2008).
81. A. Subramanian *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550 (2005).
82. J. Rameseder *et al.*, A Multivariate Computational Method to Analyze High-Content RNAi Screening Data. *Journal of Biomolecular Screening* **20**, 985-997 (2015).
83. R. Marcotte *et al.*, Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells. *Cancer Discovery* **2**, 172-189 (2012).
84. D. D. Shao *et al.*, ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Research* **23**, 665-678 (2013).
85. F. Schmich *et al.*, gespeR: a statistical model for deconvoluting off-target-confounded RNA interference screens. *Genome Biology* **16**, 1-12 (2015).
86. D. M. Garcia *et al.*, Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol* **18**, 1139-1146 (2011).
87. A. Tsherniak *et al.*, Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e516.
88. J.-P. Gillet, S. Varma, M. M. Gottesman, The Clinical Relevance of Cancer Cell Lines. *JNCI: Journal of the National Cancer Institute* **105**, 452-458 (2013).
89. S. V. Sharma, D. A. Haber, J. Settleman, Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer* **10**, 241-253 (2010).
90. R. H. Shoemaker, The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813 (2006).
91. U. T. Shankavaram *et al.*, CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics* **10**, 277 (2009).
92. A. Goodspeed, L. M. Heiser, J. W. Gray, J. C. Costello, Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Mol Cancer Res* **14**, 3-13 (2016).
93. D. Chen, M. Frezza, S. Schmitt, J. Kanwar, Q. P. Dou, Bortezomib as the First Proteasome Inhibitor Anticancer Drug: Current Status and Future Perspectives. *Current Cancer Drug Targets* **11**, 239-253 (2011).
94. J. Adams *et al.*, Proteasome Inhibitors: A Novel Class of Potent and Effective Antitumor Agents. *Cancer Research* **59**, 2615 (1999).
95. U. Scherf *et al.*, A gene expression database for the molecular pharmacology of cancer. **24**, 236 (2000).

96. O. D. Abaan *et al.*, The Exomes of the NCI-60 Panel: A Genomic Resource for Cancer Biology and Systems Pharmacology. *Cancer Research* **73**, 4372 (2013).
97. J. Gandhi *et al.*, Alterations in Genes of the EGFR Signaling Pathway and Their Relationship to EGFR Tyrosine Kinase Inhibitor Sensitivity in Lung Cancer Cell Lines. *PLoS ONne* **4**, e4576 (2009).
98. B. R. Zeeberg *et al.*, Concordance of Gene Expression and Functional Correlation Patterns across the NCI-60 Cell Lines and the Cancer Genome Atlas Glioblastoma Samples. *PLoS One* **7**, e40062 (2012).
99. M. Lukk *et al.*, A global map of human gene expression. **28**, 322 (2010).
100. D. T. Ross *et al.*, Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* **24**, 227 (2000).
101. U. McDermott *et al.*, Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proceedings of the National Academy of Sciences* **104**, 19936-19941 (2007).
102. H. W. Cheung *et al.*, Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci U S A* **108**, 12372-12377 (2011).
103. B. Luo *et al.*, Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci U S A* **105**, 20380-20385 (2008).
104. J. L. Wilding, W. F. Bodmer, Cancer Cell Lines for Drug Discovery and Development. *Cancer Research* **74**, 2377 (2014).
105. K. M. Tveit, A. Pihl, Do cell lines in vitro reflect the properties of the tumours of origin? A study of lines derived from human melanoma xenografts. *Br J Cancer* **44**, 775-786 (1981).
106. A. V. Roschke *et al.*, Karyotypic Complexity of the NCI-60 Drug-Screening Panel. *Cancer Research* **63**, 8634 (2003).
107. V. C. Daniel *et al.*, A Primary Xenograft Model of Small-Cell Lung Cancer Reveals Irreversible Changes in Gene Expression Imposed by Culture *Cancer Research* **69**, 3364 (2009).
108. W. Yang *et al.*, Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research* **41**, D955-D961 (2013).
109. C. Klijn *et al.*, A comprehensive transcriptional portrait of human cancer cell lines. **33**, 306 (2014).
110. R. M. Neve *et al.*, A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515-527 (2006).
111. H. Li *et al.*, Genomic Analysis of Head and Neck Squamous Cell Carcinoma Cell Lines and Human Tumors: A Rational Approach to Preclinical Model Selection. *Molecular Cancer Research* **12**, 571 (2014).
112. D. Mouradov *et al.*, Colorectal Cancer Cell Lines Are Representative Models of the Main Molecular Subtypes of Primary Cancer. *Cancer Research* **74**, 3238 (2014).

113. S. Domcke, R. Sinha, D. A. Levine, C. Sander, N. Schultz, Evaluating cell lines as tumour models by comparison of genomic profiles. **4**, 2126 (2013).
114. W. M. Lin *et al.*, Modeling Genomic Diversity and Tumor Dependency in Malignant Melanoma. *Cancer Research* **68**, 664 (2008).
115. M. L. Sos *et al.*, Predicting drug susceptibility of non–small cell lung cancers based on genetic lesions. *The Journal of Clinical Investigation* **119**, 1727-1740 (2009).
116. J. H. Taube *et al.*, Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences* **107**, 15449-15454 (2010).
117. J. M. Pérez-Pomares, R. Muñoz-Chápuli, Epithelial–mesenchymal transitions: A mesodermal cell strategy for evolutive innovation in Metazoans. *The Anatomical Record* **268**, 343-351 (2002).
118. J. P. Thiery, H. Acloque, R. Y. J. Huang, M. A. Nieto, Epithelial-Mesenchymal Transitions in Development and Disease. *Cell* **139**, 871-890 (2009).
119. S. A. Mani *et al.*, The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* **133**, 704-715 (2008).
120. J. Yang, S. A. Mani, R. A. Weinberg, Exploring a New Twist on Tumor Metastasis. *Cancer Research* **66**, 4549 (2006).
121. M. Oft, R. J. Akhurst, A. Balmain, Metastasis is driven by sequential elevation of H-ras and Smad2 levels. *Nat Cell Biol* **4**, 487 (2002).
122. J. Comijn *et al.*, The Two-Handed E Box Binding Zinc Finger Protein SIP1 Downregulates E-Cadherin and Induces Invasion. *Molecular Cell* **7**, 1267-1278 (2001).
123. G. Z. Cheng *et al.*, Twist Transcriptionally Up-regulates AKT2 in Breast Cancer Cells Leading to Increased Migration, Invasion, and Resistance to Paclitaxel. *Cancer Research* **67**, 1979 (2007).
124. A. D. Yang *et al.*, Chronic Oxaliplatin Resistance Induces Epithelial-to-Mesenchymal Transition in Colorectal Cancer Cell Lines. *Clinical Cancer Research* **12**, 4147 (2006).
125. Q.-Q. Li *et al.*, Twist1-Mediated Adriamycin-Induced Epithelial-Mesenchymal Transition Relates to Multidrug Resistance and Invasive Potential in Breast Cancer Cells. *Clinical Cancer Research* **15**, 2657 (2009).
126. A. Singh, J. Settleman, EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. *Oncogene* **29**, 4741-4751 (2010).
127. K. Polyak, R. A. Weinberg, Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat. Rev. Cancer* **9**, 265 (2009).
128. X. Li *et al.*, Intrinsic Resistance of Tumorigenic Breast Cancer Cells to Chemotherapy. *JNCI: Journal of the National Cancer Institute* **100**, 672-679 (2008).
129. M. Diehn *et al.*, Association of reactive oxygen species levels and radioresistance in cancer stem cells. *Nature* **458**, 780-783 (2009).

130. C. E. Eyler, J. N. Rich, Survival of the Fittest: Cancer Stem Cells in Therapeutic Resistance and Angiogenesis. *Journal of Clinical Oncology* **26**, 2839-2845 (2008).
131. E. Charafe-Jauffret *et al.*, Breast Cancer Cell Lines Contain Functional Cancer Stem Cells with Metastatic Capacity and a Distinct Molecular Signature. *Cancer Research* **69**, 1302 (2009).
132. C. M. Fillmore, C. Kuperwasser, Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. *Breast Cancer Research* **10**, R25 (2008).
133. Piyush B. Gupta *et al.*, Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. *Cell* **146**, 633-644 (2011).
134. R. A. Mathis, E. S. Sokol, P. B. Gupta, Cancer cells exhibit clonal diversity in phenotypic plasticity. *Open Biology* **7**, (2017).
135. T. Kondo, T. Setoguchi, T. Taga, Persistence of a small subpopulation of cancer stem-like cells in the C6 glioma cell line. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 781-786 (2004).
136. L. J. Harper, K. Piper, J. Common, F. Fortune, I. C. Mackenzie, Stem cell patterns in cell lines derived from head and neck squamous cell carcinoma. *Journal of Oral Pathology & Medicine* **36**, 594-603 (2007).
137. R. M. Meyers *et al.*, Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet*, (2017).
138. T. Wang *et al.*, Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell* **168**, 890-903 e815 (2017).
139. K. Tzelepis *et al.*, A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep* **17**, 1193-1205 (2016).
140. T. Hart *et al.*, High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515-1526 (2015).
141. T. Wang *et al.*, Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101 (2015).
142. P. B. Gupta *et al.*, Identification of Selective Inhibitors of Cancer Stem Cells by High-Throughput Screening. *Cell* **138**, 645-659 (2009).
143. T. Pemovska *et al.*, Axitinib effectively inhibits BCR-ABL1(T315I) with a distinct binding conformation. *Nature* **519**, 102-105 (2015).
144. J. W. Tyner *et al.*, Kinase Pathway Dependence in Primary Human Leukemias Determined by Rapid Inhibitor Screening. *Cancer Research* **73**, 285 (2013).
145. C. Hatzis *et al.*, Enhancing reproducibility in cancer drug screening: how do we move forward? *Cancer Res* **74**, 4016-4023 (2014).
146. P. Geeleher, E. R. Gamazon, C. Seoighe, N. J. Cox, R. S. Huang, Consistency in large pharmacogenomic studies. *Nature* **540**, E1-E2 (2016).

147. M. Bouhaddou *et al.*, Drug response consistency in CCLE and CGP. *Nature* **540**, E9-E10 (2016).
148. J. P. Mpindi *et al.*, Consistency in drug response profiling. *Nature* **540**, E5-E6 (2016).
149. P. M. Haverty *et al.*, Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333-337 (2016).
150. T. Dobzhansky, Genetics of natural populations. XIII. Recombination and variability in populations of Drosophila pseudoobscura. *Genetics* **31**, 269 (1946).
151. J. L. Hartman, B. Garvik, L. Hartwell, Principles for the Buffering of Genetic Variation. *Science* **291**, 1001 (2001).
152. L. H. Hartwell, P. Szankasi, C. J. Roberts, A. W. Murray, S. H. Friend, Integrating Genetic Approaches into the Discovery of Anticancer Drugs. *Science* **278**, 1064 (1997).
153. S. H. Friend, A. Oliff, Emerging Uses for Genomic Information in Drug Discovery. *New England Journal of Medicine* **338**, 125-126 (1998).
154. W. G. Kaelin Jr, The Concept of Synthetic Lethality in the Context of Anticancer Therapy. *Nat. Rev. Cancer* **5**, 689 (2005).
155. S. M. B. Nijman, Synthetic lethality: General principles, utility and detection using genetic screens in human cells. *FEBS Letters* **585**, 1-6 (2011).
156. H. Farmer *et al.*, Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917-921 (2005).
157. P. C. Fong *et al.*, Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from BRCA Mutation Carriers. *New England Journal of Medicine* **361**, 123-134 (2009).
158. H. E. Bryant *et al.*, Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913-917 (2005).
159. R. L. Beijersbergen, L. F. A. Wessels, R. Bernards, Synthetic Lethality in Cancer Therapeutics. *Annual Review of Cancer Biology* **1**, 141-161 (2017).
160. J. Downward, RAS Synthetic Lethal Screens Revisited: Still Seeking the Elusive Prize? *Clinical Cancer Research* **21**, 1802 (2015).
161. N. J. O'Neil, M. L. Bailey, P. Hieter, Synthetic lethality and cancer. *Nat. Rev. Genet.* **18**, 613 (2017).
162. L. Jerby-Arnon *et al.*, Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality. *Cell* **158**, 1199-1209 (2014).
163. M. W. Libbrecht, W. S. Noble, Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16**, 321 (2015).
164. Machine learning: An artificial intelligence approach. RS Michalski, JG Carbonell, TM Mitchell (Eds), Berlin Heidelberg GmbH: Springer-Verlag; 1983.
165. F. Azuaje, Computational models for predicting drug responses in cancer research. *Briefings in Bioinformatics* **18**, 820-829 (2017).
166. S. Okser *et al.*, Regularized Machine Learning in the Genetic Prediction of Complex Traits. *PLOS Genetics* **10**, e1004754 (2014).

167. I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, A. A. Margolin, Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac Symp Biocomput*, 63-74 (2014).
168. M. Gönen, A Bayesian Multiple Kernel Learning Framework for Single and Multiple Output Regression. *In Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012), Montpellier, France*, (2012).
169. J. C. Costello *et al.*, A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202 (2014).
170. T. Hofmann, B. Schölkopf, A. J. Smola, Kernel Methods in Machine Learning. *The Annals of Statistics* **36**, 1171-1220 (2008).
171. K. J. Bussey *et al.*, Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Molecular Cancer Therapeutics* **5**, 853 (2006).
172. A. Daemen *et al.*, Modeling precision treatment of breast cancer. *Genome Biology* **14**, R110 (2013).
173. M. Ammad-Ud-Din, S. A. Khan, K. Wennerberg, T. Aittokallio, Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. *Bioinformatics* **33**, i359-i368 (2017).
174. M. Ammad-Ud-Din *et al.*, Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* **32**, i455-i463 (2016).
175. Z. Safikhani *et al.*, Gene isoforms as expression-based biomarkers predictive of drug response in vitro. *Nat Commun* **8**, 1126 (2017).
176. J. W. Kim *et al.*, Characterizing genomic alterations in cancer by complementary functional associations. *Nat. Biotechnol.* **34**, 539 (2016).
177. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. *Springer-Verlag New York*, (2009).
178. M. Vidal, M. E. Cusick, A. L. Barabasi, Interactome networks and human disease. *Cell* **144**, 986-998 (2011).
179. Y. Zhu, P. Qiu, Y. Ji, TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods* **11**, 599-600 (2014).
180. L. Kong *et al.*, ESTOOLS Data@Hand: human stem cell gene expression resource. *Nat Methods* **10**, 814-815 (2013).
181. P. Naula, A. Airola, T. Salakoski, T. Pahikkala, Multi-label learning under feature extraction budgets. *Pattern Recognition Letters* **40**, 56-65 (2014).
182. T. Pahikkala, S. Okser, A. Airola, T. Salakoski, T. Aittokallio, Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms Mol Biol* **7**, 11 (2012).
183. http://dreamchallenges.org/.
184. J. Saez-Rodriguez *et al.*, Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet* **17**, 470-486 (2016).

185. R. Silberzahn, E. L. Uhlmann, Crowdsourced research: Many hands make tight work. *Nature* **526**, 189-191 (2015).
186. H. W. Cheung *et al.*, Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci U S A* **108**, (2011).
187. R. Marcotte *et al.*, Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov* **2**, 172-189 (2012).
188. F. D. Sigoillot *et al.*, A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. *Nat Methods* **9**, (2012).
189. B. Yilmazel *et al.*, Online GESS: prediction of miRNA-like off-target effects in large-scale RNAi screen data by seed region analysis. *BMC Bioinformatics* **15**, 192 (2014).
190. E. Buehler *et al.*, siRNA off-target effects in genome-wide screens identify signaling pathway members. *Sci rep* **2**, (2012).
191. R. Zhong *et al.*, Computational detection and suppression of sequence-specific off-target phenotypes from whole genome RNAi screens. *Nucleic Acids Research* **42**, 8214-8222 (2014).
192. O. Shalem *et al.*, Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343**, 84-87 (2014).
193. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **343**, 80-84 (2014).
194. T. Hart *et al.*, High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515-1526 (2015).
195. J. G. Doench *et al.*, Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotech* **34**, 184-191 (2016).
196. B. Evers *et al.*, CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat Biotech* **34**, 631-633 (2016).
197. D. W. Morgens, R. M. Deans, A. Li, M. C. Bassik, Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat Biotech* **34**, 634-636 (2016).
198. Y. Fu *et al.*, High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotech* **31**, 822-826 (2013).
199. P. D. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotech* **31**, 827-832 (2013).
200. J. G. Doench *et al.*, Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotech* **32**, 1262-1267 (2014).
201. R. E. Neapolitan, X. Jiang, Study of integrated heterogeneous data reveals prognostic power of gene expression for breast cancer survival. *PLoS One* **10**, e0117658 (2015).
202. A. A. Margolin *et al.*, Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci Transl Med* **5**, 181re181 (2013).

203. F. Eduati *et al.*, Prediction of human population responses to toxic compounds by a collaborative competition. *Nat Biotech* **33**, 933-940 (2015).

204. K. Eppert *et al.*, Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat Med* **17**, 1086-1093 (2011).

Recent Publications in this Series

14/2018 Jukka Saarinen
Non-linear Label-free Optical Imaging of Cells, Nanocrystal Cellular Uptake and Solid-State Analysis in Pharmaceutics
15/2018 Olena Santangeli
Sleep and Depression: Developmental and Molecular Mechanisms
16/2018 Shadia Rask
Diversity and Health in the Population: Findings on Russian, Somali and Kurdish Origin Populations in Finland
17/2018 Richa Gupta
Association and Interplay of Genetic and Epigenetic Variants in Smoking Behavior
18/2018 Patrick Vingadas Almeida
Multifunctional Porous Silicon Based Nanocomposites for Cancer Targeting and Drug Delivery
19/2018 Lena Sjöberg
Reproductive Health in Women with Childhood-onset Type 1 Diabetes in Finland
20/2018 Perttu Päiviö Salo
Studies on the Genetics of Heart Failure
21/2018 Andrew Erickson
In Search of Improved Outcome Prediction of Prostate Cancer – A Biological and Clinical Approach
22/2018 Imrul Faisal
Genetic Regulation of Mammalian Spermatogenesis - Studies of USF1 and MAD2
23/2018 Katja Wikström
Socioeconomic Differences in the Development and Prevention of Type 2 Diabetes: Focus on Education and Lifestyle
24/2018 Laura Ollila
Genotype-Phenotype Correlations in Dilated Cardiomyopathy
25/2018 Elina Engberg
Physical Activity, Pregnancy and Mental Wellbeing: Focus on Women at Risk for Gestational Diabetes
26/2018 Anni Niskakoski
Molecular Alterations of Endometrial and Ovarian Tumorigenesis in Lynch Syndrome Mutation Carriers and the General Population
27/2018 Katariina Maaninka
Atheroinflammatory Properties of LDL and HDL Particles Modified by Human Mast Cell Neutral Proteases
28/2018 Sonja Paetau
Neuronal ICAM-5 Regulates Synaptic Maturation and Microglia Functions
29/2018 Niina Kaartinen
Carbohydrates in the Diet of Finnish Adults - Focus on Intake Assessment and Associations with Other Dietary Components and Obesity
30/2018 Tuija Jääskeläinen
Public Health Importance of Vitamin D: Results from the Population-based Health 2000/2011 Survey
31/2018 Tiina Lipiäinen
Stability and Analysis of Solid-State Forms in Pharmaceutical Powders
32/2018 Johanna Ruohoalho
Complications and Their Registration in Otorhinolaryngology – Head and Neck Surgery: Special emphasis in Tonsil Surgery Quality Registration

Helsinki 2018 **ISSN 2342-3161** **ISBN 978-951-51-4294-8**