Integration of Multiple Data Types for Analysis of Cancer Cell Lines

Dmitriy Sonkin

A submission presented in partial fulfillment of the

requirements of the University of South Wales / Prifysgol De Cymru

for the degree of Doctor of Philosophy

December 11th 2013

University of South Wales

Table of Contents

CCLE_MSigDB.c1.all.v3.0.entrez_GeneSets_PermutationFraction.txt

(MSigDB v3.0 Cytogenetic Bands Gene Sets Permutation Fractions in CCLE)

CCLE_MSigDB.c2.cgp.v3.0.entrez_GeneSets_ActivityScores.txt

(MSigDB v3.0 Genetic and Chemical Perturbations Gene Sets Activity Scores in CCLE)

CCLE_MSigDB.c2.cgp.v3.0.entrez_GeneSets_PermutationFraction.txt

(MSigDB v3.0 Genetic and Chemical Perturbations Gene Sets Permutation Fractions in CCLE)

CCLE_MSigDB.c2.cp.v3.0.entrez_GeneSets_ActivityScores.txt

(MSigDB v3.0 Canonical Pathways Activity Scores in CCLE)

CCLE_MSigDB.c2.cp.v3.0.entrez_GeneSets_PermutationFraction.txt

(MSigDB v3.0 Canonical Pathways Permutation Fractions in CCLE)

CCLE_MSigDB.c3.mir.v3.0.entrez_GeneSets_ActivityScores.txt

(MSigDB v3.0 microRNA targets Gene Sets Activity Scores in CCLE)

CCLE_MSigDB.c3.mir.v3.0.entrez_GeneSets_PermutationFraction.txt

(MSigDB v3.0 microRNA targets Gene Sets Permutation Fractions in CCLE)

CCLE_MSigDB.c3.tft.v3.0.entrez_GeneSets_ActivityScores.txt

(MSigDB v3.0 Transcription Factor Targets Gene Sets Activity Scores in CCLE)

CCLE_MSigDB.c3.tft.v3.0.entrez_GeneSets_PermutationFraction.txt

(MSigDB v3.0 Transcription Factor Targets Gene Sets Permutation Fractions in CCLE)

Abstract

Cancer cell lines play an important and critical part in oncology research. The advances in understanding of cancer biology which were achieved in the last few decades would be virtually impossible without using cancer cell lines as research models. Therefore better understanding of molecular properties of such models is crucial element in cancer research. Recently collaborations between Sanger Institute and Massachusetts General Hospital Cancer Center and also between Broad institute and Novartis Institutes for BioMedical Research Inc. generated mRNA expression, copy number, microRNA expression, sequencing and compound sensitivity data for each of the cell lines from the collection covering almost a thousand of the available cancer cell lines. Such data provides rich sources to explore important insights into tumor biology; however they also highlight the need for additional approaches for integrative analysis of multiple data types. The work presented in this thesis is significant contribution to the efforts to make use of all available data per sample and existing biological knowledge to the greatest possible extent. In particular this work covers two research topics: tumor suppressor genes status and gene sets activity analysis on sample by sample basis.

Tumor suppressor genes play a major role in the etiology of human cancer, and typically achieve a tumor promoting effect upon complete functional inactivation. Bi-allelic inactivation of tumor suppressors may occur through genetic mechanisms (such as loss-of-function mutations, DNA loss), epigenetic mechanisms (such as promoter methylation or histones modifications), signaling mechanisms or a combination of these inactivation mechanisms. Prior to the work presented in this thesis no nomenclature system existed in order to capture the complexity of tumor suppressor genes functional status and correspondingly no computational framework existed to generate such status. In order to address this deficiency, a comprehensive nomenclature system and computational framework was developed for the assessment of tumor suppressor genes functional “status”. It is utilizing several orthogonal genomic data types, including mutation data, copy number, LOH and expression. Through correlation with additional data types (compound sensitivity and gene set activity) it is shown that this integrative method, which allows accounting for multiple mechanisms of tumor suppressor genes inactivation, provides a more accurate assessment of tumor suppressor genes status than can be inferred by expression, copy number, or mutation alone. The utilization of this comprehensive and systematic computational framework led to marked improvement in annotation of TP53 status across extensive collection of cancer cell lines. Identifying cell lines with high confidence wild type TP53 status provides critically important foundation for efforts to identify signature to predict sensitivity to inhibitors of MDM2 driven degradation of TP53.

Approach to perform gene set activity on sample by sample basis is covered in this thesis along with its application to the extensive collection of the cancer cell lines. Underlying implementation is used in part to establish pSTAT5 mRNA expression signature in hematopoietic cancer cell lines. This signature can potentially make it possible to identify patients whom may benefit from JAK inhibitor(s), based on JAK-STAT signaling.

Acknowledgements

This Ph.D. study has its roots going back in time and across borders of different countries and also across the walls of a number of institutions and organizations. I had to mentally traverse the route which brought me to this point in order to at least try to thank the key people whose help was instrumental in my education.

I thank Dr. Tatiana V. Tatarinova for guiding me through the Ph.D. journey providing invaluable support and encouragements through the whole endeavor. Her knowledge and ability to find time to talk across multiple time zones was extremely helpful and in fact absolutely essential. I thank Professor Denis J. Murphy for help with navigating the world of scientific publishing and guiding me through the administrative steps, which sometimes were more complicated than usual due to changes to departments and institutions. I thank my supervisor Dr. Michael Morrissey at Novartis Institutes for BioMedical Research for making it possible for me to pursue my Ph.D. study. Without his knowledge, encouragements and incredible patience this work would not be possible.

Also I need to thank Dr. Heidi L. Rehm, Dr. Birgit H. Funke and Dr. Victoria Joshi, with whom I had the privilege to work at Harvard Partners Center for Genetics and Genomics, for helping me to better understand human genetics and its application to bio-medical research and patient care.

Also I need to thank Dr. Viktor M. Dyadkin who was supervisor of my master degree thesis, his help was instrumental in starting my professional career. And of course I need to thank all the teachers who help me at one point or another during the last few decades.

Author’s Declaration

I declare that the work in this thesis was carried out in accordance with the requirements of the University’s Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate’s own work. Work done in collaboration with, or with the assistance of others, is indicated as such. Any views expressed in the thesis are those of the author.

SIGNED Sorkis **DATE** December 11th 2013

List of Figures

List of Tables

List of Abbreviations and Definitions

ABL1	c-abl oncogene 1
AML	Acute Myeloid Leukemia
ALL	Acute Lymphoblastic Leukemia
BAP1	BRCA1 Associated Protein 1
BCR	Breakpoint Cluster Region
bp	Base Pair(s)
CCLE	Cancer Cell Line Encyclopedia
CML	Chronic Myeloid Leukemia
CN	Copy Number
CNS	Central Nervous System
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic Acid
DSB	Double Strand Break
FAP	Familial Adenomatous Polyposis
FDR	False Discovery Rate
IC50	Concentration at which the drug response reached an absolute inhibition of 50%

IGV	Integrative Genomics Viewer
JAK	Janus tyrosine Kinase
kbp	kilo bp
LFL	Li-Fraumeni-like syndrome
LFS	Li-Fraumeni syndrome
LOH	Loss Of Heterozygosity
MAPK	Mitogen Activated Protein Kinase
MEFs	Mouse Embryonic Fibroblasts
mRNA	Messenger RNA
MRT	Malignant Rhabdoid Tumors
NCI	National Cancer Institute
PCA	Principal component analysis
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
SSB	Single Strand Break
STAT	Signal Transducer and Activator of Transcription
TCGA	The Cancer Genome Atlas

Chapter One: Introduction

This thesis focuses on two main research topics: tumor suppressor genes status and gene sets activity analysis on sample by sample basis.

Tumor suppressor genes play a major role in the etiology of human cancer, and they typically have a tumor promoting effect upon complete functional inactivation. Bi-allelic inactivation of tumor suppressors may occur through genetic mechanisms (such as loss-of-function mutations, DNA loss), epigenetic mechanisms (such as promoter methylation or histones modifications), signaling mechanisms or a combination of these inactivation mechanisms. Currently there is no nomenclature system in order to capture the complexity of the functional status of tumor suppressor genes. There is no computational framework to generate such status. In order to address this deficiency, a comprehensive nomenclature system and computational framework are developed in this thesis. These components are essential for the assessment of the functional "status" of tumor suppressor genes based on several orthogonal genomic data types, such as mutation data, copy number, LOH and expression. In this thesis, the developed framework is used to generate tumor suppressor genes status for extensive collection of cancer cell lines.

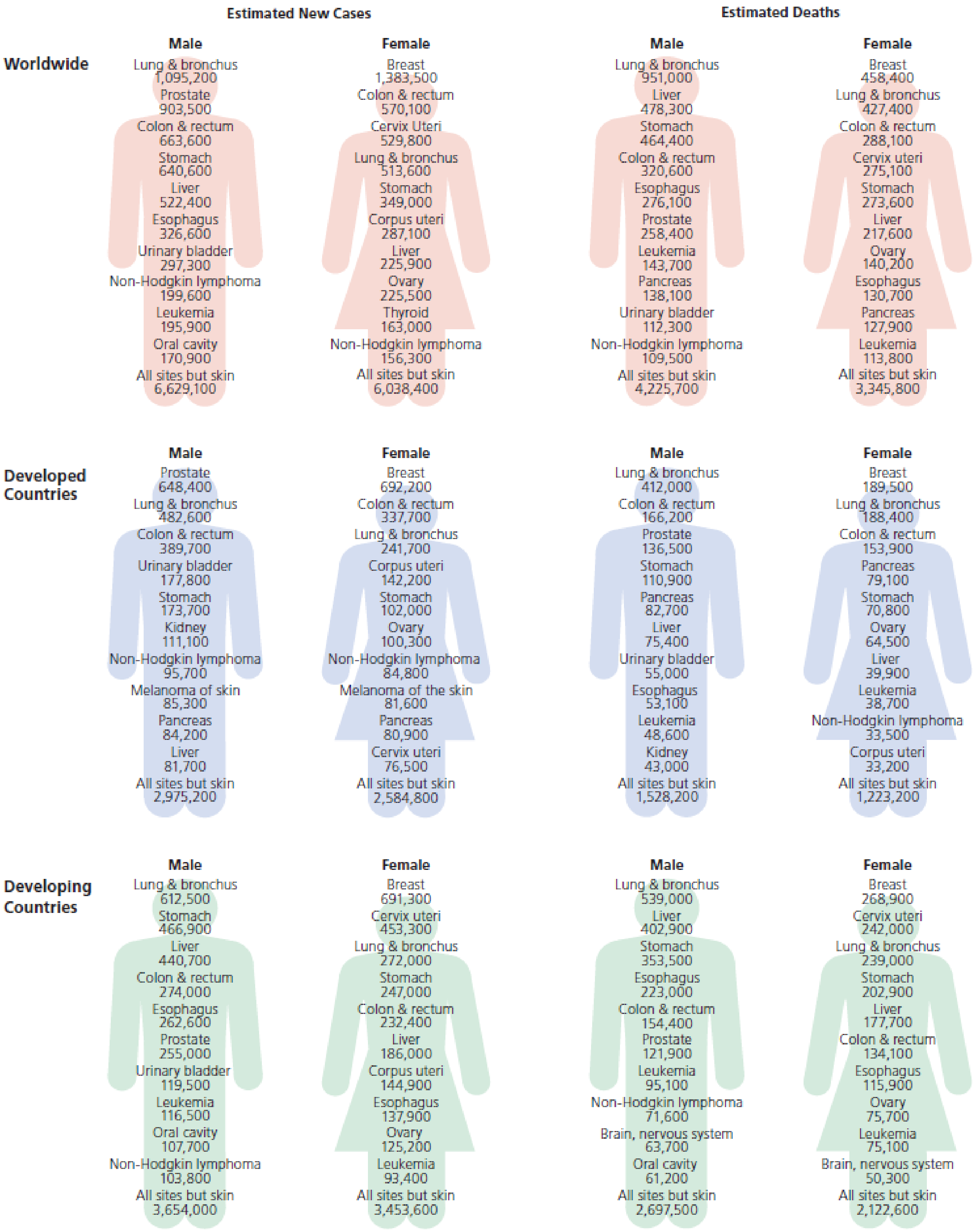
Here is a brief description of the content. Chapter 1.1 provides background information on cancer cell lines as models for cancer research and describes collection of cancer cell lines used for this work. Chapter 1.2 provides background information on tumor suppressor genes biology and their critical importance in cancer biology.

It is often desirable to obtain measures of gene sets activity on a sample-by-sample basis. The computational framework developed in this thesis offers a *z-score* based implementation to generate gene sets activity on a sample-by-sample basis and integrates it with permutation based method to access statistical significance. Chapter 1.3 provides background information on gene sets activity analysis research. Later in Chapter 1.3, gene sets activity analysis on a sample-by-sample basis is used in part to establish pSTAT5 mRNA expression signature in hematopoietic cancer cell lines. Chapter 1.4 provides background information on biology of JAK-STAT signaling. Chapter 1.5 outlines the key research steps to develop comprehensive computational frameworks to (a) access status of tumor suppressor genes and (b) generate gene sets activity on a sample-by-sample basis.

1.1 Cancer cell lines

Cancer cell lines play an important and critical part in oncology research. The advances in understanding of cancer biology which were achieved in the last few decades would be virtually impossible without using cancer cell lines as research models.

Cancer is one of the leading causes of death in the world. The Figure 1.1 shows the global cancer statistics from 2011, indicating that cancers of various types affect millions of people each year around the world (Jemal et al., 2011). Clearly there is need for more pre-clinical, translational and clinical research in order to decrease the burden of malignant diseases on individuals and society at large. It is important to note that most dramatic improvements in cancer treatment are often due to improvements in understanding of cancer biology.



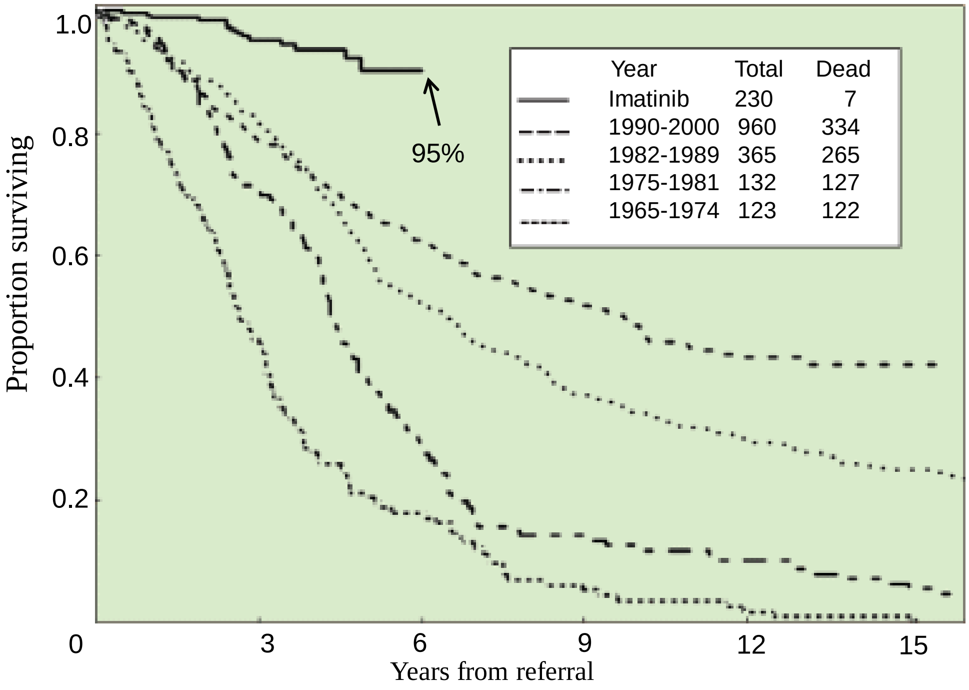
A. Jemal et al. A Cancer Journal for Clinicians, 2011, volume 61(2):69-90 (Source: GLOBOCAN 2008)

Figure 1.1 Global cancer statistics

The classical and important example of such improvement is a case of Chronic Myeloid Leukemia (CML). The chromosomal abnormality driving CML was discovered in 1959 and became known as “Philadelphia Chromosome" (Nowell and Hungerford, 1961). In 1973 it was shown that “Philadelphia Chromosome" is a reciprocal translocation between chromosomes 9 and 22 (Rowley, 1973).After a decade of the research efforts it was shown that translocation between chromosomes 9 and 22 results in fusion of Breakpoint Cluster Region (BCR) gene and c-abl oncogene 1 (ABL1) gene (Groffen et al., 1984). In 2001 clinical trials demonstrated that ABL1 inhibitor imatinib (brand name: Gleevec) is an extremely effective and well tolerated treatment for CML (Druker et al., 2001). It is important to note that the development of ABL1 inhibitors did not stop at that point, since about 25-35 percent of CML patients do not achieve complete cytogenetic remission or become eventually resistant to imatinib (Shah et al., 2002). The second generation of ABL1 inhibitors such as dasatinib (brand name: Sprycel) (Shah et al., 2004) and nilotinib (brand name: Tasigna) (Weisberg et al., 2005) are effective against all but one known ABL1 mutant (T315I). Third generation ABL1 inhibitor ponatinib (brand name: Iclusig) is currently in clinical trials and initial results seems to indicate activity against T315I ABL1 mutant (Cortes et al., 2012). It is possible to imagine a scenario in the future where CML patients are potentially treated with combination of different ABL1 inhibitors in order to greatly reduce chance of encountering resistance to treatment.

In just few years after introduction of ABL1 inhibitor(s) into clinical practice it became very clear that prognosis for CML patients have been drastically improved. Figure 1.2 demonstrates this fact looking at survival of CML patients treated at M. D. Anderson Cancer Center over several decades (Quintás-Cardama and Cortes, 2006). A few years later a multicenter study showed that about 95% of CML patients continued to have complete cytogenetic remission after

6 years of treatment and about 90% of CML patients continued to have complete cytogenetic remission after 8 years of treatment (Gambacorti-Passerini et al., 2011). Remarkably in that study deaths due to CML accounted for only 1% of CML patients after 8 years of treatment and overall life expectancy of CML patients was almost the same as a matched age group of the general population. Worldwide there are estimated 100,000 new cases of CML each year (Jemal et al., 2011), indicating that lives of hundreds of thousands CML patients have been saved due to advances in CML treatment.



Quintás-Cardama A, Cortes JE, Mayo Clinic Proceedings Volume 81, Issue 7 2006 973 - 988

Figure 1.2 Survival of patients with CML treated at M. D. Anderson Cancer Center

The progress in CML treatment would be very unlikely without usage of CML cancer cell lines derived from patients with CML. For example, the work in K562 CML cancer cell line led to discovery of tyrosine kinase activity of BCR-ABL fusion (Konopka et al., 1984). Also CML cancer cell lines with various ABL1 mutations leading to resistance to imatinib were instrumental in developing new generations of ABL1 inhibitors.

The first cancer cell line was successfully isolated from a patient with aggressive cervical cancer and grown *in vitro* in 1951(Gey et al., 1952). This cell line known as HeLa played an important role in overall biological research and not just oncology research. For example this cell line played an important role in work on polio vaccine (Scherer et al., 1953). In the following decades hundreds of other cancer cell lines were established and made available for researchers around the world. The National Cancer Institute (NCI) established the panel of 60 different human cancer cell lines for testing anti-cancer compounds, this collection became known as NCI-60 (Shoemaker et al., 1988). The NCI-60 panel became the first extensively used panel of human cancer cell lines for high throughput compound testing (Monks et al., 1991) (Weinstein et al., 1997). More than 60,000 compounds have been tested against the NCI-60 panel.

The cancer cell lines have been generated from malignancies covering majority of cancer types (Garnett et al., 2012) (Barretina et al., 2012). However there is a wide range of success in deriving cancer cell lines from primary tissue samples. For some indications such as breast carcinomas and melanomas the success rate is relatively high, while for others such as for example prostate cancers the success rate is rather low and only a limited number of unique cell lines is available for research (Sobel and Sadar, 2005). Also sometimes it is rather difficult to generate a cancer cell line from the primary tumor driven by particular alteration, for example astrocytoma tumors with EGFR amplifications (Pandita et al., 2004). The difficulties are due in part to complex and sometimes unclear growth conditions required by some of the tumor types which may for example depend on the presence of particular ligand that may not be readily available in typical animal based cell media (Scheithauer et al., 1987). Recently published work on use of ROCK inhibitor and feeder cells may be important step in increasing spectra of primary tumors from which cancer cell lines may be established (Liu, Ory, et al., 2012).

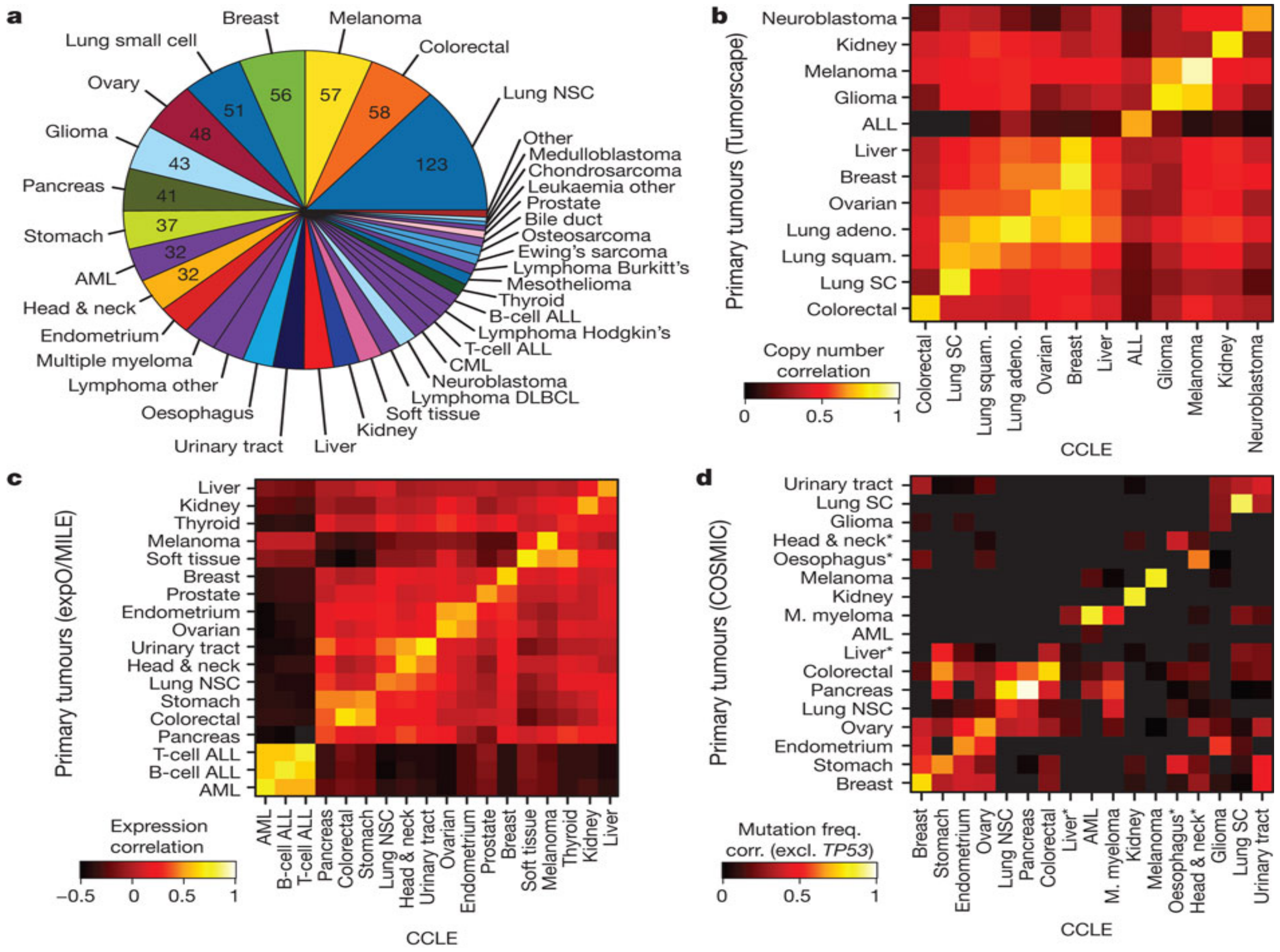
Patient-derived tumor xenografts are another type of cancer model which has received significant attention in last few years (Tentler et al., 2012). It is interesting to note that tumor xenografts may be the better starting point for cancer cell lines generation than primary tumors (Dangles-Marie et al., 2007).

The role of the cancer microenvironment is a growing area of cancer research (Ungefroren et al., 2011). The research in this area lead to realization of the importance of *in vivo* environment and therefore the difference in behavior between cancer cells and primary tumors (Gillet et al., 2013). One of the potential ways to address possible differences between *in vivo* and *in vitro* growth conditions is to grow cell lines in three dimensional scaffolds (Nyga et al., 2011).

The Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) provides a comprehensive molecular characterization of nearly a thousand cancer cell lines and therefore lays out a foundation for better understanding of cancer cell lines biology. The CCLE provides mRNA expression, Affymetrix SNP 6.0 profiles, OncoMap (MacConaill et al., 2009) mutation screening and exome sequencing data. This rich data set allows a number of different analyses using a combination of multiple data types. This multiplicity allows this research project to examine distinct mechanisms of tumor suppressor inactivation across CCLE cell lines.

The following four figures from (Barretina et al., 2012) provide a general overview of CCLE: Figure 1.3a shows distribution of cancer types in the CCLE by lineage. This figure shows that CCLE covers in reasonable depth a majority of main tumor types. Figure 1.3b shows that there is reasonable correlation between CCLE CN profiles and primary tumors CN profiles using as a measure GISTIC G-scores (Mermel et al., 2011) across CCLE and 12 corresponding cancer types included in Tumorscape (Beroukhim et al., 2010). Figure 1.3c shows that there is reasonable correlation between CCLE mRNA expression profiles and 18 corresponding primary tumors cancer types from the expO (http://www.intgen.org/expo) and MILE datasets (Haferlach et al., 2010). Figure 1.3d shows that there is reasonable correlation between an oncogene point mutation in CCLE cell lines and corresponding primary tumors in COSMIC (Forbes et al., 2011) for 378 mutations in 29 genes.

Affymetrix U133Plus2 mRNA expression, Affymetrix SNP 6.0 data, OncoMap mutation calls (MacConaill et al., 2009), exome data sequencing (Hodges et al., 2007), and pharmacological profiling data are available at the CCLE web site (http://www.broadinstitute.org/ccle/home). Expression data is MAS5 normalized, with a 2% trimmed mean of 150 (Hubbell et al., 2002).



J Barretina *et al*. *Nature* 483, 603-607 (2012) doi:10.1038/nature11003

Figure 1.3a Distribution of cancer types in the CCLE by lineage.

Figure 1.3b CCLE CN profiles vs. primary tumors CN profiles.

The diagonal of the heat map shows the Pearson correlation between corresponding tumor types.

Figure 1.3c CCLE mRNA profiles vs. primary tumors mRNA profiles.

For each tumor type, the log fold change of the 5,000 most variable genes is calculated between that tumor type and all others. Pearson correlations between tumor type fold changes from primary tumors and cell lines are shown as a heat map.

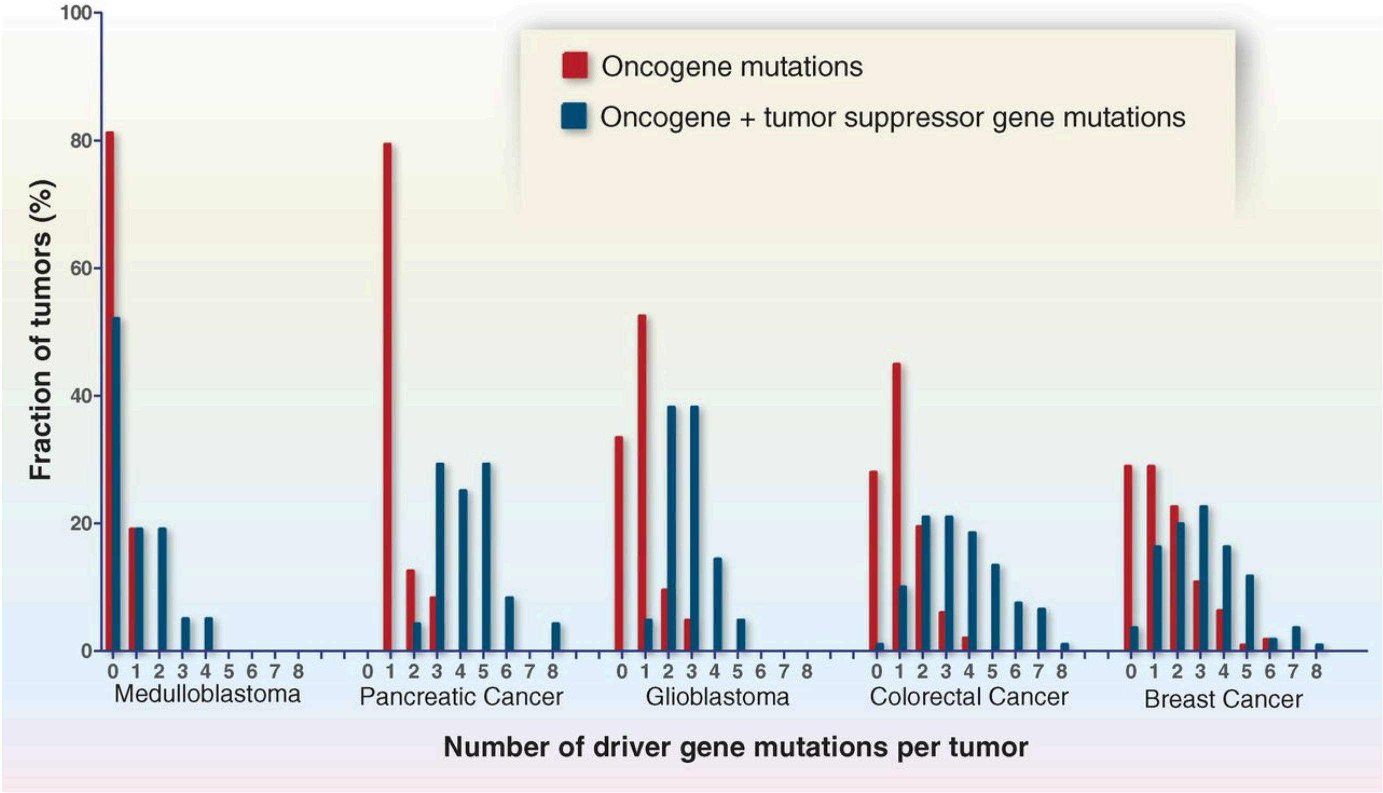
Figure 1.3d CCLE point mutations vs. primary tumors point mutations.

Comparison of point mutation frequencies between cell lines and primary tumors in COSMIC (v56), restricted to genes that are well represented in both sample sets but excluding TP53, which is highly prevalent in most tumors types. Pairwise Pearson correlations are shown as a heat map. Asterisk indicates that the correlations of oesophageal, liver, and head and neck cancer mutation frequencies are restored when including TP53.

1.2 Tumor suppressor genes status

Tumor suppressor genes encode proteins that normally inhibit tumor formation caused by abnormal cellular proliferation. Tumor suppressor proteins can participate in a variety of processes such as negative regulation of the cell cycle, positive regulation of apoptosis, regulation of DNA damage response, or other mechanisms (Stanbridge, 1990). The list of tumor suppressor genes includes such names as TP53 (tumor protein p53), RB1 (retinoblastoma 1), APC (adenomatous polyposis coli), and BRCA1 (breast cancer 1, early onset). The inactivation of these and other tumor suppressor genes plays a major role in many types of cancer (Jones and Thompson, 2009). Also, in general, tumor suppressor genes inactivation may be even more frequent events than oncogene activations per individual tumor in most solid malignancies, as can be seen in Figure 1.4 from (Vogelstein et al., 2013).

The first tumor suppressor gene, RB1, was proposed in 1971 (Knudson, 1971). Knudson’s elegant work suggested that both copies of RB1 need to be inactivated in one way or another in order for a tumor to form and therefore two hits are needed in order to disable a tumor suppressor gene. Knudson’s work was based on detailed examination of two types of retinoblastomas: familial and sporadic. The familial form of retinoblastomas appears (as their name suggests) in families with a history of disease and sporadic retinoblastomas arises in children with families without history of disease. The familial form of retinoblastomas leads to retinoblastomas affecting both eyes (bilateral) while sporadic retinoblastomas affect only a single eye (unilateral).



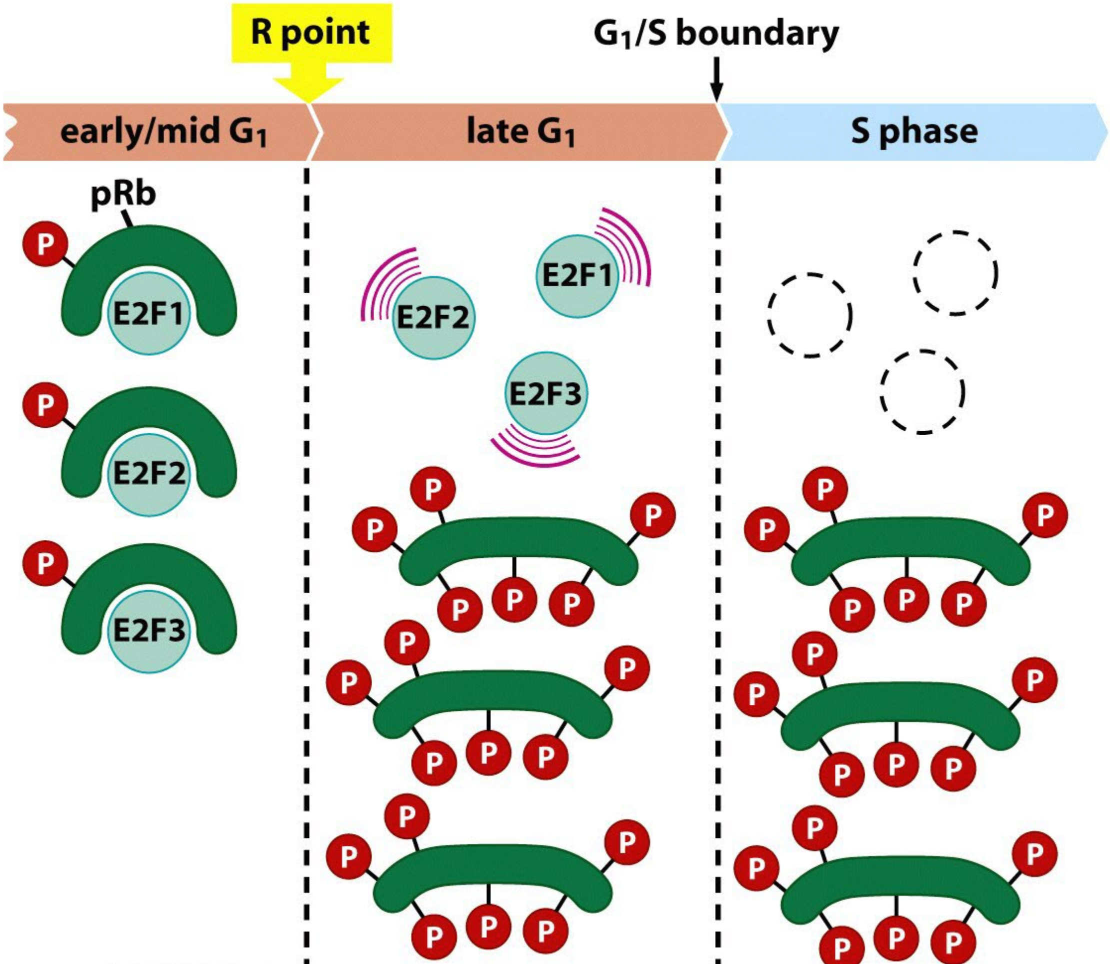
Vogelstein et al. Science 2013;339:1546-1558

Figure 1.4 Number and distribution of driver gene mutations in five tumor types.

The total number of driver gene mutations [in oncogenes and tumor suppressor genes (TSGs)] is shown, as well as the number of oncogene mutations alone.

Bilateral retinoblastomas appear early in life in comparison to unilateral forms. The Knudson hypothesis was based on examination of the kinetics of bilateral and unilateral retinoblastomas. Knudson proposed that in cases of familial retinoblastoma one copy of RB1 was inherited in mutant form and therefore only one additional hit to RB1 was required in order to generate a tumor. In the case of sporadic retinoblastoma, two somatic inactivating events to RB1 are required in order to form the tumor. In 1986 RB1 was cloned and as predicted there were inactivating alterations affecting both copies of the gene in affected patients (Friend et al., 1986). Children with the familial form of retinoblastoma have at least 500 times above normal increased risk of developing osteosarcoma tumors during their life time (Kleinerman et al., 2005). Further research determined that RB1 plays a critical role in cell cycle control (Chellappan et al., 1991), (Nevins, 2001). Figure 1.5 shows the relationship between RB1, E2Fs and cell cycle (Weinberg, 2007). Figure 1.6 shows RB1 and other tumor suppressor genes and oncogenes involved in regulation of restriction point transition (Weinberg, 2007).

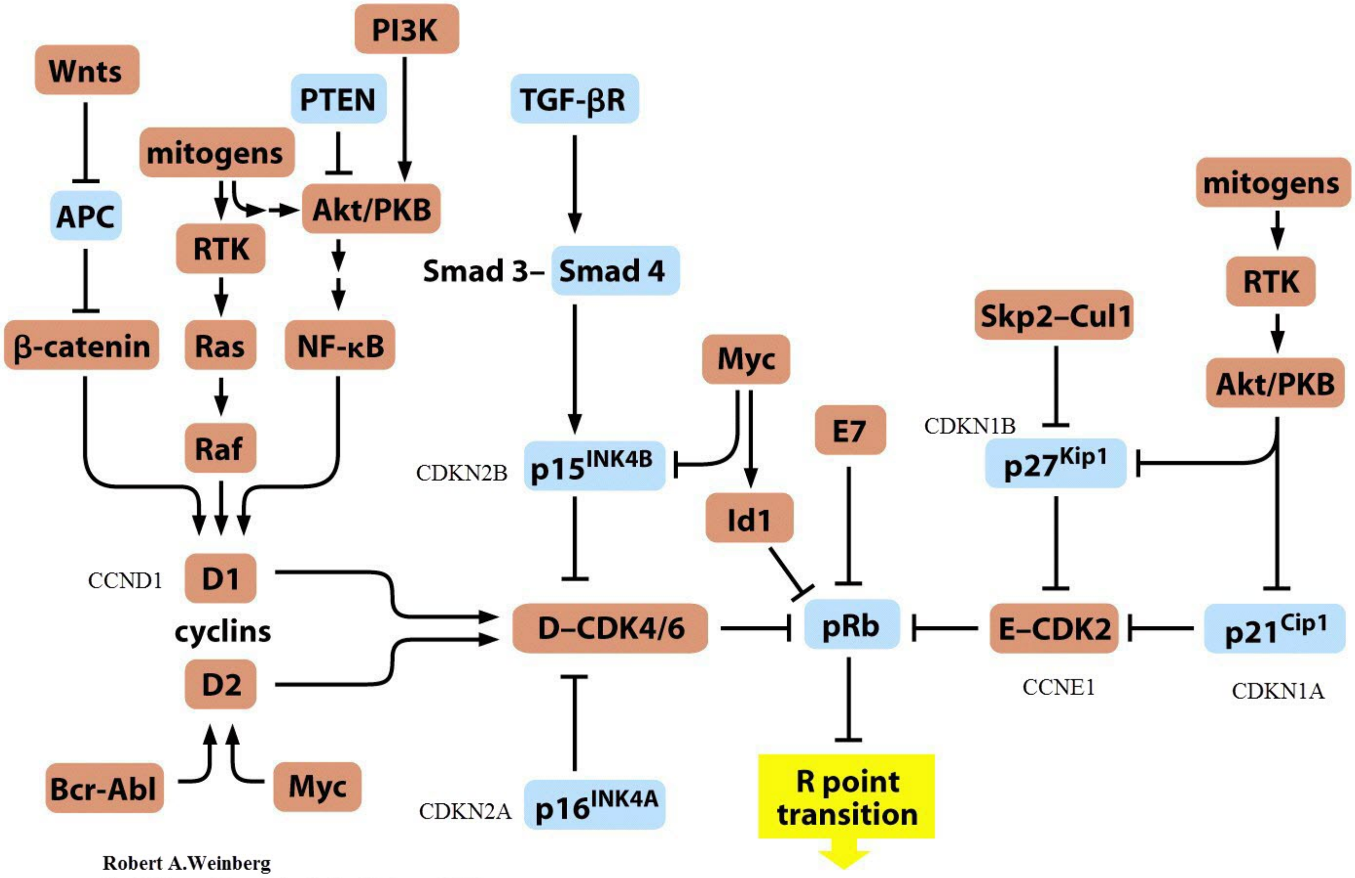
It is important to keep in mind that retinoblastoma tumors belong to a rather small, but quite interesting group of malignancies in which a loss of just one tumor suppressor gene leads to cancer (Goodrich, 2006). Another example of such tumor type is malignant rhabdoid tumors (MRT), which are rare, mostly pediatric, tumors of kidney, liver, soft tissue and central nervous system (CNS) (Wick et al., 1995). In MRT loss of SNF5 is sufficient to cause tumor formation (McKenna et al., 2008), in fact loss of SNF5 is a critical diagnostic marker. It is interesting to note that SNF5 is a core component of the SWI/SNF chromatin remodeling complex (Muchardt and Yaniv, 1999). In recent years number of genes such as ARID1A, ARID1B, SMARCA2, SMARCA4, etc. encoding subunits of chromatin remodeling complexes have been found to be inactivated in a number of different malignances and with noticeable frequency (Wu and Roberts, 2013) (Shain and Pollack, 2013).



Robert A.Weinberg
The Biology of Cancer (© Garland Science 2007)

Figure 1.5 RB1, E2Fs and cell cycle

Hypophosphorylated RB1 blocks transcription activating domain of E2Fs. Hyperhosphorylated RB1 releases E2Fs allowing them to activate transcription of genes required for progression through cell cycle. E2Fs are inactivated and/or degraded as cells enter S phase.



Robert A.Weinberg
The Biology of Cancer (© Garland Science 2007)

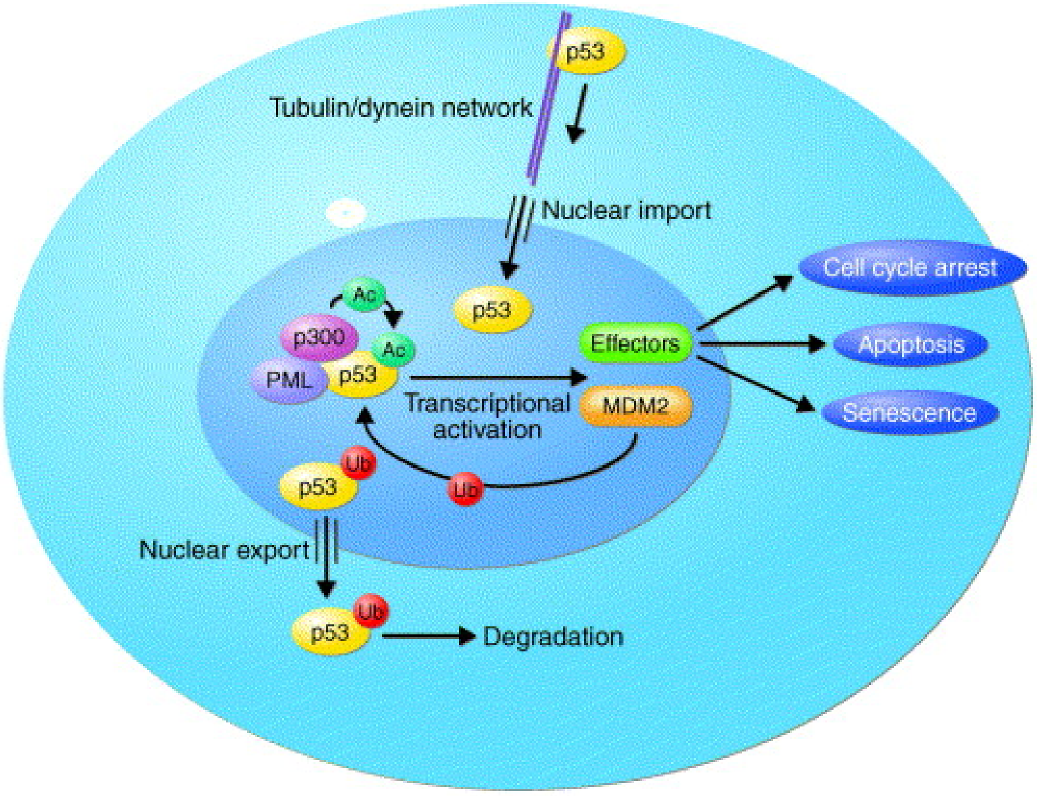
Figure 1.6 Restriction point transition signaling

Elements that promote advance through the R point are drawn in orange, while those that block this advance are shown in blue.

Both retinoblastoma and malignant rhabdoid tumors are pediatric cancers, which in comparison to adult tumors in general have smaller number of genetic alterations (Vogelstein et al., 2013). A recent paper investigating Acute lymphoblastic leukemia (ALL) cancers in two pairs of identical twins ages 55 and 48 months respectively, showed the same pattern of rather low number of genetic alterations (Ma et al., 2013). Relatively small numbers of genetic alterations are not completely exclusive to pediatric cancers, for example acute myeloid leukemia (AML) and chronic lymphocytic leukemia (CLL) have relatively few genetic alterations (Vogelstein et al., 2013). The Cancer Genome Atlas (TCGA) project is massive effort to get detailed molecular characterization of thousands primary tumors covering a number of different indications, this project is helping to get a more detailed picture of genetic alterations in cancer (Cancer Genome Atlas Research Network, 2008).

At this point there are about 80 well-known and putative tumor suppressor genes. Numbers of them are inactivated with rather high frequency. For example TP53 is one of the most frequently mutated genes in cancer (Petitjean et al., 2007). The functional status of number of tumor suppressor genes directly influences the selection of possible treatment options. For example CDK4/6 inhibitors can only be potentially effective against tumors with wild type RB1 (Finn et al., 2009). Also inhibitors of MDM2-driven TP53 protein degradation can only be potentially effective against tumors with wild type TP53 (Efeyan et al., 2007). Figure 1.7 shows a simplified diagram of regulation and function of the TP53 (Ryan et al., 2001). On the other hand PARP inhibitors are more likely to be effective in tumors with inactivated BRCA1 or BRCA2, since PARP enzymes are involved in DNA repair and their inhibition can lead to

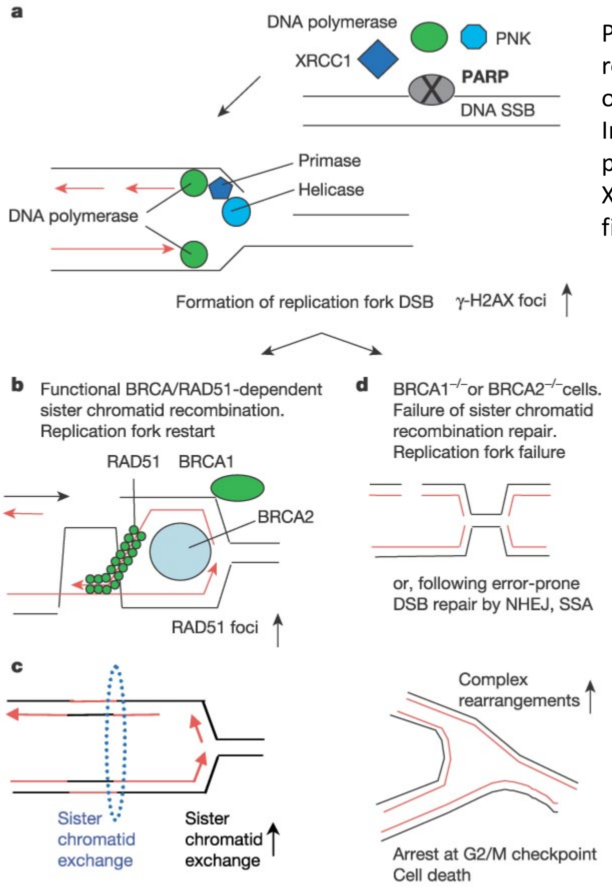
DNA breaks and BRCA1 and BRCA2 are involved in homologous repair of DNA breaks (Kummar et al., 2012).



Current Opinion in Cell Biology

Ryan KM, Phillips AC, Vousden KH. 2001 Jun;13(3):332-7.

Figure 1.7 Regulation and function of the TP53


a
DNA polymerase
PNK
XRCC1
PARP
DNA SSB
Primase
Helicase
DNA polymerase
Formation of replication fork DSB γ-H2AX foci ↑
b
Functional BRCA/RAD51-dependent sister chromatid recombination. Replication fork restart
RAD51
BRCA1
BRCA2
RAD51 foci ↑
c
Sister chromatid exchange
Sister chromatid exchange ↑
d
BRCA1$^{-/-}$or BRCA2$^{-/-}$cells. Failure of sister chromatid recombination repair. Replication fork failure
or, following error-prone DSB repair by NHEJ, SSA
Complex rearrangements ↑
Arrest at G2/M checkpoint Cell death

PARP functions in base excision repair. DNA SSBs form due to oxidative damage and its repair. Inhibition of PARP activity prevents the recruitment of XRCC1 and subsequent SSB gap filling by DNA polymerases.

Large numbers of DNA SSBs persist and are encountered by DNA replication forks. These lead to replication fork arrest associated with a DSB.

Presence of functional BRCA1 and BRCA2, allows initiation of sister chromatid recombination repair. A collapsed replication fork may be restarted by this mechanism.

When Holliday junctions at recombination intermediates are resolved, a sister chromatid exchange may occur. The excess number of replication fork arrests associated with loss of PARP function leads to an increase in sister chromatid recombination events and sister chromatid exchanges.

In the absence of functional BRCA1 or BRCA2, sister chromatid recombination and the formation of RAD51 foci are severely impaired. Replication-associated DSBs cannot be repaired by sister chromatid recombination. Some remain unrepaired as chromatid breaks but many are repaired by error-prone RAD51-independent mechanisms such as non-homologous end joining (NHEJ) and single-strand annealing (SSA).

Farmer et al. Nature 434, 917-921 (14 April 2005)

Figure 1.8 PARP inhibition and BRCA1/2 status

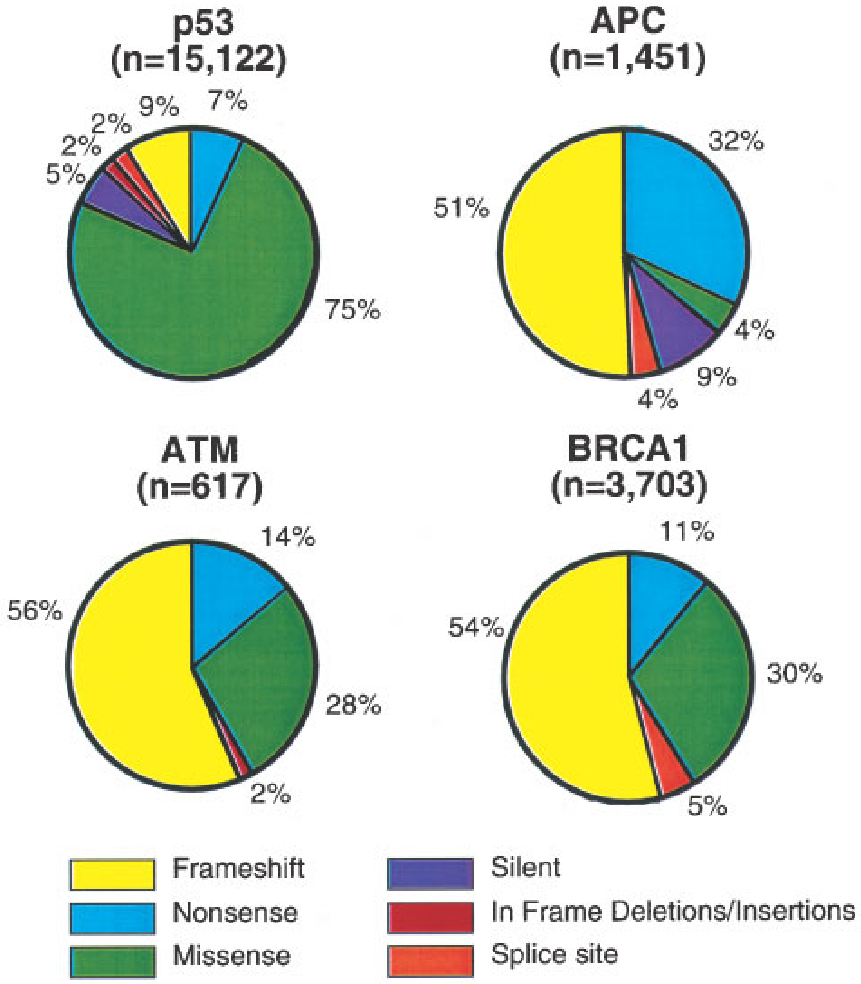
Figure 1.8 illustrates consequences of PARP inhibition in cell with functioning BRCA1 and BRCA2 versus cell with loss of function of BRCA1 or BRCA2 (Farmer et al., 2005). Clinical progress in using PARP inhibitors was complicated by use of an iniparib compound which was supposed to be PARP inhibitor, but failed clinical trials and was eventually shown to be not a PARP inhibitor in the first place (Liu, Shi, et al., 2012) (Patel et al., 2012). Despite this setback clinical development of PARP inhibitors continued and some of them are now in Phase 3 of clinical trials. Table 1.1 lists PARP inhibitors in clinical trials along with information on clinical phase and indication(s) (Source http://clinicaltrials.gov). Phase 3 clinical trials for inhibitors niraparib and olaparib as part of the trials are going to evaluate the impact of BRCA1 and BRCA2 status in relation to therapeutic respond to the compounds.

Table 1.1 List of PARP inhibitors in clinical trials.

Compound	Sponsor	Phase	Indication(s)
Niraparib	Tesaro	Phase 3	Ovarian cancer
Olaparib (AZD-2281)	AstraZeneca	Phase 3	BRCA1/2 mutant cancers, solid tumors
Veliparib (ABT-888)	Abbott	Phase 2	Prostate, colorectal, leukemia, solid tumors
CEP-9722	Cephalon	Phase 2	Solid tumors, lymphoma
Rucaparib (CO-338)	Clovis Oncology	Phase 2	BRCA1/2 mutant cancers, solid tumors
E7016	Eisai	Phase 2	Solid tumors
BMN-673	BioMarin Pharmaceutical	Phase 1	Leukemia, solid tumors

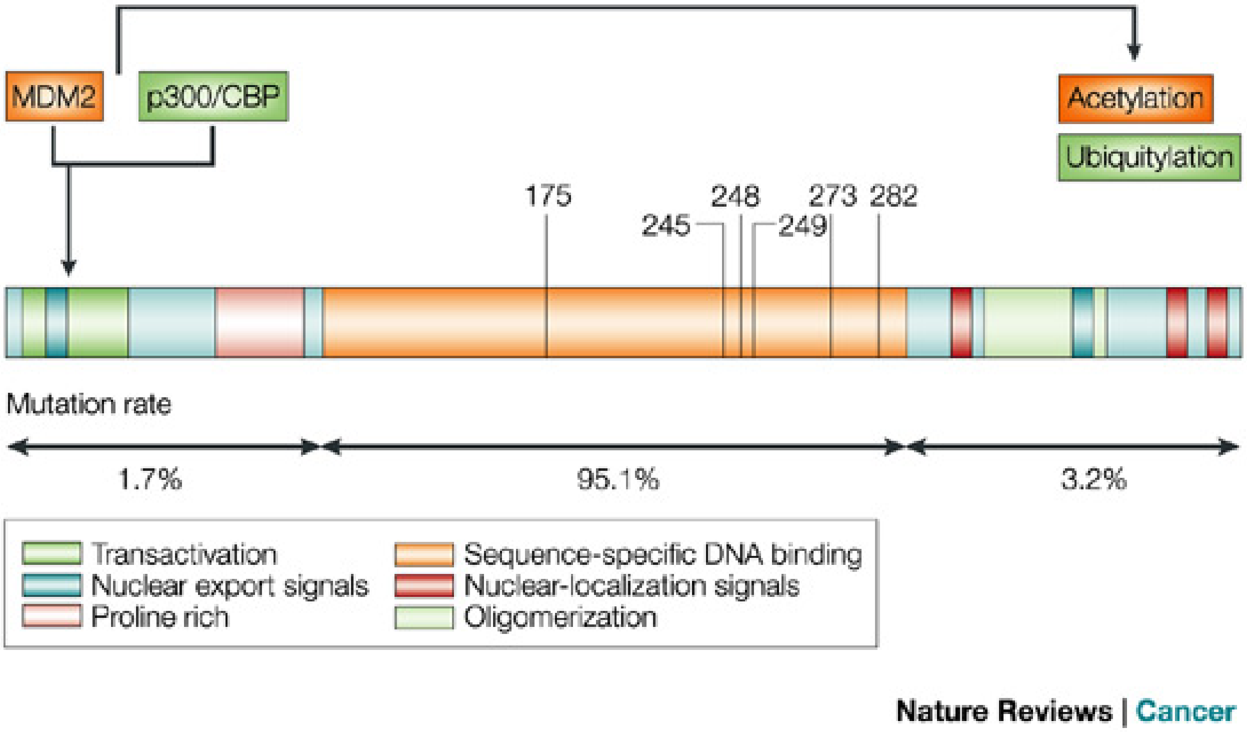
Figure 1.9 shows classes of mutations found in the tumor suppressor genes TP53, APC, ATM and BRCA1 (Robles et al., 2002). APC, ATM and BRCA1 have a classical pattern of mutations found in the tumor suppressor genes with majority of mutations belonging to nonsense or frame-shift classes. However TP53 displays a very different mutation pattern with majority of mutations belonging to a missense class. Figure 1.10 provides a detailed view of the TP53 mutation patterns (Vousden and Lu, 2002). The main reason for such unusual mutational patterns for a tumor suppressor gene is related to the fact that TP53 needs to form a tetramer in order to perform its transcriptional factor function. Therefore even if only one out of four subunits has a missense mutation it may behave in a dominant negative fashion and therefore prevent correct formation of the tetramer (Petitjean et al., 2007). To this end, six highlighted residues on Figure 1.10 represent locations of TP53 hotspot mutations. Mutations in these six residues account for about 28% of all TP53 mutations and non-surprisingly almost all mutations at these locations are considered to be dominant negative ones.

As previously mentioned unlike proto-oncogenes (Croce, 2008), where a single mutation can be dominant and lead to cellular transformation, a single mutation in a tumor suppressor gene is normally recessive as long as there is a second functional copy of the gene. However, loss of function of both tumor suppressor alleles may promote tumor growth or survival providing that the loss-of-function is nearly or totally complete. It is possible to infer loss-of-function of tumor suppressor genes through a number of genomic measurements, such as mRNA transcript expression, DNA copy numbers, and sequencing data.



Robles AI, Linke SP, Harris CC Oncogene. 2002 Oct 7;21(45):6898-907

Figure 1.9 Mutations found in the tumor suppressor genes TP53, APC, ATM, BRCA1



Vousden KH and Lu X. 2002 Aug;2(8):594-604

Figure 1.10 Mutations pattern in the tumor suppressor gene TP53

Some particular types of alterations leading to inactivation of tumor suppressors may not necessary be frequent events, but still are interesting case studies. For example DNA deletions which do not affect the coding part of the genes, but instead remove promoter or essential enhancer sequences, and also translocations which move promoter or essential enhancer sequences away from coding part of the gene; as a result both of these alterations may lead to loss of expression of tumor suppressor genes. The above mentioned mechanisms for inactivation of tumor suppressors can be broadly divided in three categories.

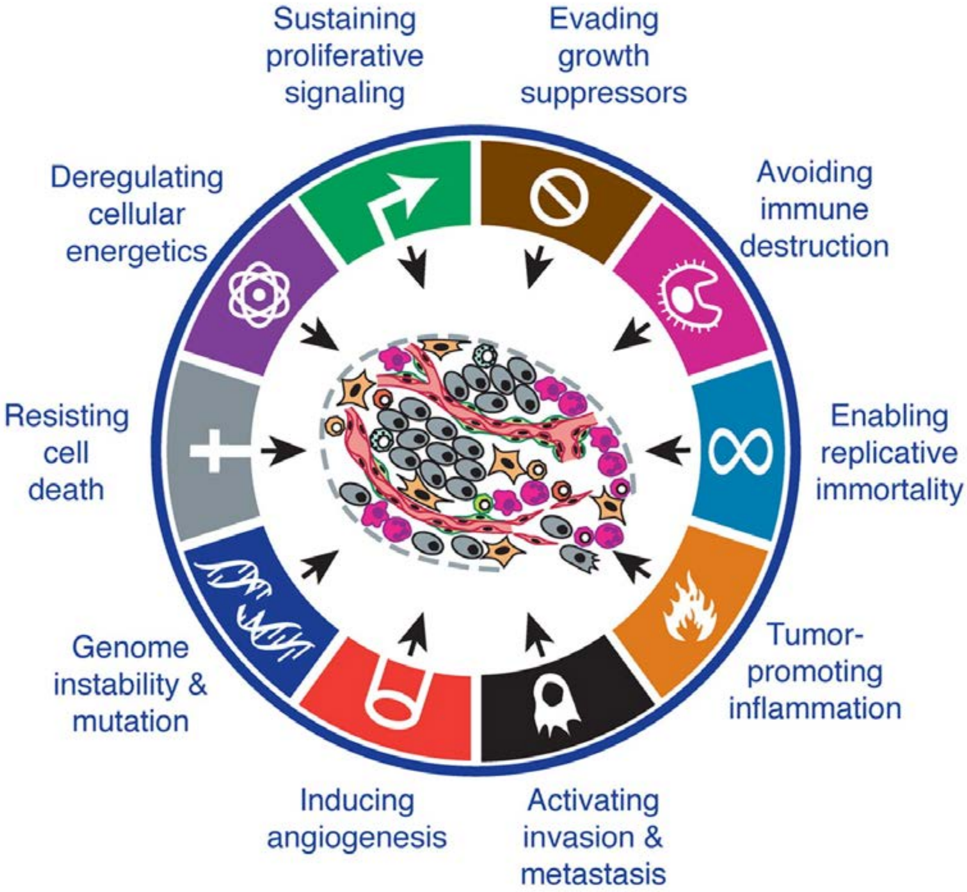
The first category includes inactivation of both alleles by genetic alterations, such as copy number loss, loss of heterozygosity (LOH) and mutations.

The second category includes inactivation of one allele by a mechanism from the first category and loss of mRNA expression of second allele by an epigenetic mechanism, such as promoter methylation, possible histone modifications and other mechanisms leading to loss of mRNA expression.

The third category includes inactivation of both alleles by an epigenetic mechanism.

The comprehensive and systematic computational framework is presented in chapter two allowing examination of tumor suppressor for gene and sample in question and assignment of appropriate status.

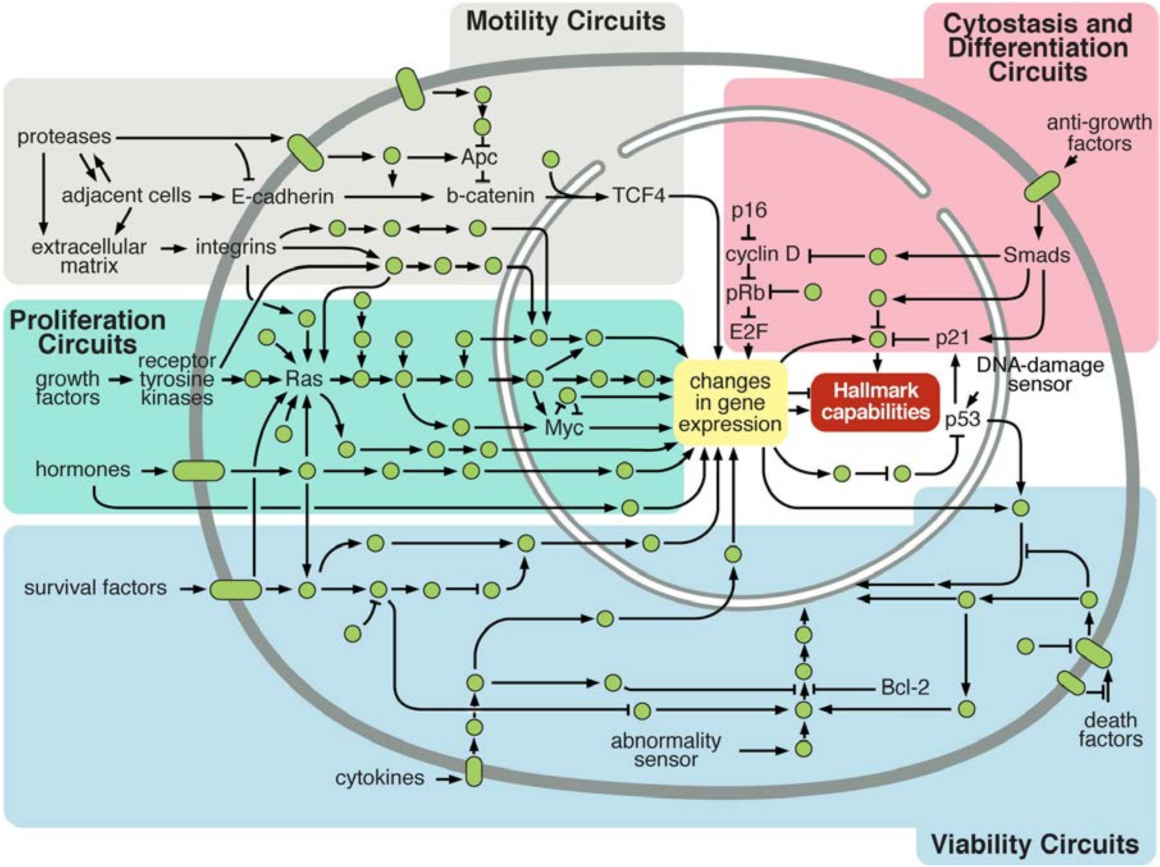
It is important to keep in mind that tumorigenesis is a complicated and multifaceted process. Hallmarks of cancer are the key conceptual characteristics likely needed for successful tumorigenesis and they are depicted in Figure 1.11 (Hanahan and Weinberg, 2011). There are number of underlying molecular mechanisms behind these hallmarks of cancer.


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

Douglas Hanahan , Robert A. Weinberg Cell Volume 144, Issue 5 2011 646 - 674

Figure 1.11 Hallmarks of cancer

Figure 1.12 shows a schematic and simplified map of known intracellular signaling networks which are critical components of tumorigenesis (Hanahan and Weinberg, 2011). Gene sets activity analysis introduced in Chapter 1.3 is an approach that uses existing knowledge of intracellular signaling networks in order to better understand complicated processes in individual tumors.



Douglas Hanahan , Robert A. Weinberg Cell Volume 144, Issue 5 2011 646 - 674

Figure 1.12 Intracellular signaling networks in cancer

1.3 Gene sets activity analysis

Gene expression analysis on the whole genome level is an important technique in molecular biology (Schena et al., 1995) (Lockhart et al., 1996) (Lee et al., 2008). There are number of different approaches for analyzing gene expression data sets. Most of the initial bioinformatics work on gene expression analysis on the whole genome level was concerned with choosing appropriate test statistics to identify differentially expressed genes (Tusher et al., 2001) (Smyth, 2004) and on the applying multiple hypothesis corrections (Hochberg and Benjamini, 1990) (Storey and Tibshirani, 2003) (Dudoit et al., 2003). One group of approaches for analyzing gene expression data is trying to take advantage of existing knowledge of molecular pathways available from resources such as Gene Ontology (Ashburner et al., 2000), MSigDB (Liberzon et al., 2011), (Kanehisa et al., 2012), BIOCARTA (http://www.biocarta.com) and GeneGo (www.genego.com). Pathway activity analysis methods can be classified into two major groups: over representation and the aggregate score approaches.

One of widely known approaches from aggregate score category is generally known as Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003) (Subramanian et al., 2005). There are numerous ways to calculate aggregate scores and multiple publications have explored the different approaches (Pavlidis et al., 2002) (Goeman et al., 2004) (Kim and Volsky, 2005) (Tian et al., 2005) (Dinu et al., 2007) (Irizarry et al., 2009) (Hänzelmann et al., 2013).

The signal to noise ratio is used by the GSEA method in order to rank gene expression differences between two groups of samples.

$S_i = \frac{E(x_{ij}|j \in Y_1) - E(x_{ij}|j \in Y_2)}{\sigma^t(x_{ij}|j \in Y_1) + \sigma^t(x_{ij}|j \in Y_2)}$, where

S_i is a signal to noise ratio for the gene i.

x_{ij} is an expression value for the gene *i* in the sample j.

Y_1 is a first set of samples.

Y_2 is a second set of samples.

$E(x_{ij}|j \in Y_1)$ is an expected value for the first set of samples.

$E(x_{ij}|j \in Y_2)$ is an expected value for the second set of samples.

Expected value is also known as a mean, which for discrete random variable X_i is calculated as:

$E(X_i) = \frac{1}{N}\sum_{j=1}^{N} x_{ij}$, where *N* is a number of samples in a set for which the mean is calculated.

σ^t is a truncated standard deviation for each group of samples limited from below at 20% of the corresponding means in order to decrease spikes in signal to noise ratio due to artificially low standard deviation.

$\sigma^t(x_{ij}|j \in Y_*) = \begin{cases} \sigma(x_{ij}|j \in Y_*) & if\ \sigma(x_{ij}|j \in Y_*) \geq 0.2 \times E(x_{ij}|j \in Y_*) \\ 0.2 \times E(x_{ij}|j \in Y_*) & if\ \sigma(x_{ij}|j \in Y_*) < 0.2 \times E(x_{ij}|j \in Y_*) \end{cases}$, where Y_* represents set of samples for which standard deviation is calculated. As before, *N* is a number of samples in a set for which standard deviation is calculated.

σ is a standard deviation, defined as $\sigma(x_i) \ = \ \sqrt{\frac{\sum_{j=1}^{N} \ (x_i - E(x_i))^2}{N}}$

GSEA uses modified Kolmogorov-Smirnov statistic to assess gene set enrichment. For a kth gene set, $S_k^{GSEA} \ = \ sup_{l=1\ldots N}\left(F_l^{g_k} \ - \ F_l^{\overline{g_k}}\right)$, where S_k^{GSEA} is a enrichment score for the gene set k; N is number of genes in rank gene list; *l* is a rank in the gene list.

$F_l^{g_k}$ is a cumulative distribution function calculated for genes in gene set k.

$$F_l^{g_k} \ = \frac{\sum_{h=1}^{l} S_h I_h}{\sum_{h=1}^{N} S_h I_h}$$

where S_h is signal to noise for gene h.

$$F_l^{\overline{g_k}} = \frac{\sum_{h=1}^{l}(1 - \ I_h)}{(N \ - \ n_k)}$$

$F_l^{\overline{g_k}}$ is a cumulative distribution function calculated for genes not in gene set k.

where n_k is number of genes in the gene set k.

$$I_h \ = \begin{cases} 1 & if \ h \ \in \ g_k \\ 0 & if \ h \ \in \ \overline{g_k} \end{cases}$$

GSEA enrichment scores are normalized using the expected value of the positive or negative null distribution statistic generated by sample permutation. This normalization is necessary in order to be able to compare GSEA enrichment values across gene sets of different sizes.

$$\acute{S}_k^{GSEA} \ = \begin{cases} \dfrac{S_k^{GSEA}}{E(S_k^{GSEA} | S_k^{GSEA} \ \geq 0)} & if \ S_k^{GSEA} \ \geq \ 0 \\ \dfrac{S_k^{GSEA}}{E(S_k^{GSEA} | S_k^{GSEA} \ < 0)} & if \ S_k^{GSEA} \ < \ 0 \end{cases}$$

$\acute{S}_k^{GSEA}$ is a normalized enrichment score for gene set k.

The Kolmogorov-Smirnov statistic is non-parametric and distribution free, therefore it makes no assumptions on distribution properties of underlying data set (Gibbons, 2003). The Kolmogorov-Smirnov statistic is based on analysis of cumulative distribution function, also known as empirical distribution function. This statistical approach may therefore allow for the detection of differences between distributions even if, for example, the means of underlying distributions are the same. However in general non-parametric statistics are less sensitive if the underlying data meets requirements of the particular parametric statistical test (Freedman, 2005).

Work by (Irizarry et al., 2009) indicated that z-scores can be successfully used to calculate aggregate scores of gene set enrichment in part by comparing analysis results generated using z-scores based approach and GSEA from the same datasets. In fact (Irizarry et al., 2009) analysis demonstrated examples where z-scores based approach seems to be superior to GSEA. In z-scores based approach two-sample t-test based statistics is used to estimate differences in expression for each gene between two groups of samples.

$$t_i = \frac{E(x_{ij}|j \in Y_1) - E(x_{ij}|j \in Y_2)}{\sqrt{\sigma^2(x_{ij}|j \in Y_1) + \sigma^2(x_{ij}|j \in Y_2)}}$$

t_i is a t-statistic for the gene i.

$$S_k^z = E(t_i|i \in G_k)\sqrt{n_k}$$

S_k^z is z-score for gene set k.

G_k is set of genes in gene set k.

n_k is number of genes in the gene set k.

Statistical significance of z-scores for gene sets can be estimated using assumption of standard normal distribution.

S_k^z is not able to detect changes in scales, because such changes do not cause mean shift. For example it's possible, in some scenarios, to have a pathway in which up-regulated and down regulated genes balance each other almost exactly and the balance is dynamically preserved. To account for such cases χ^2 test is used by (Irizarry et al., 2009).

$$S_k^{\chi^2} = \frac{\sum_{i \in G_k} \left(t_i - E(t_i)\right)^2 - (n_k - 1)}{2(n_k - 1)}$$

$S_k^{\chi^2}$ is χ^2 score for gene set k.

Additional research pointed to examples in which GSEA analysis seems to generate results superior to z-scores based approach (Tamayo et al., 2012). GSEA analysis and z-scores based approach seems to produce over all similar results and each method has its own pluses and minuses which in part depend on underlying datasets and gene sets in question.

There is also a significant number of publications and tools for analysis that uses overrepresentation approach (Grosu et al., 2002) (Doniger et al., 2003) (Zeeberg et al., 2003) (Dennis et al., 2003) (Al-Shahrour et al., 2004) (Zhong et al., 2004) (Zhou and Su, 2007) (Beltrame et al., 2013). Overrepresentation approach takes a look at the list of differentially expressed genes and searches for gene sets which are represented in the list more often than would be expected just by chance.

The hypergeometric test is often used to access statistical significance of overrepresentation analysis. One tailed P-value could be calculated by the hypergeometric test in the following way:

p $= \sum_{i=x}^{K} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}$, where:

N is the total number of genes available in the collection of gene sets with mRNA expression data.

M is the number of genes in the gene set of interest.

K is the number of differentially expressed genes.

x is the number of differentially expressed genes which belong to gene set of interest.

$\binom{n}{k}$ is a binomial coefficient and its value given by $\frac{n!}{k!(n-k)!}$

The methods described in this introductory chapter up to this point require performing differential gene set analysis between two groups of samples as a first step. Due to this requirement the result of the analysis is a relative measure of gene set enrichment between two groups. However in a number of cases it is desirable to obtain measures of gene sets activity on a sample-by-sample basis. Also, in case of modern database repositories with tens of thousands gene expression profiles, it becomes increasingly difficult to use pathway enrichment analysis to generate analyses covering all profiles due to the combinatorial explosion of possible grouping of samples.

Compared to classical pathway enrichment analysis between two groups of samples, pathway activity analysis on a sample-by-sample basis is a much less well studied approach, although there is still a modest body of work in that area. Methods for pathway activity analysis on a sample-by-sample basis could be roughly divided into four generic categories:

mean/median based methods (Guo et al., 2005) (Breslin et al., 2005), z-scores based methods (Levine et al., 2006) (Lee et al., 2008), sample level extensions of GSEA (Edelman et al., 2006) (Barbie et al., 2009) (Hänzelmann et al., 2013) and principal component analysis (PCA) based methods (Tomfohr et al., 2005) (Bild et al., 2006).

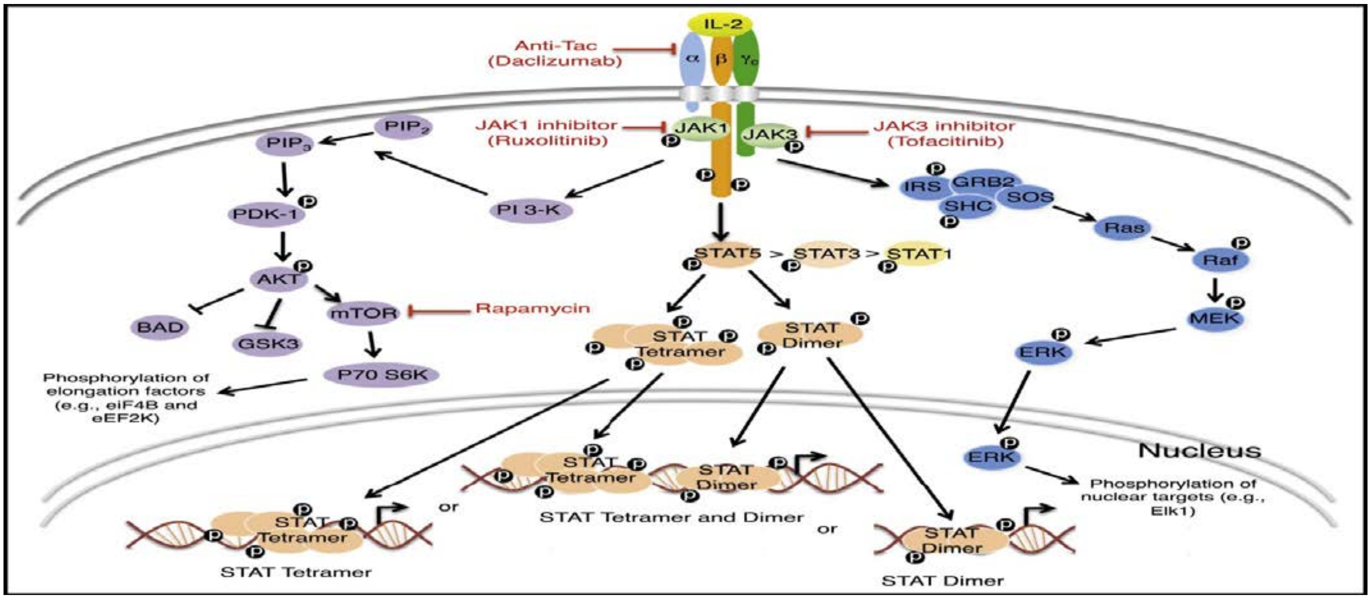
Z-score transformation is a classical method of data normalization and there is a long and solid history of using z-scores for normalization of gene expression data (Cheadle et al., 2003). Therefore it seems like the z-scores based methods for pathway activity analysis on a sample-by-sample basis are built on a firm foundation.

1.4 JAK-STAT signaling pathway

In chapter three, the application of gene sets activity analysis to Cancer Cell Line Encyclopedia cell lines is discussed and also its application to generation of pSTAT5 mRNA expression signature in hematopoietic cancer cell lines. The following paragraphs introduce the JAK-STAT pathway and its relevance to hematopoietic malignancies.

The JAK-STAT pathway is one of the key signaling pathways downstream of cytokine and growth factor receptors. It plays a critical role in hematopoiesis, immune functions and many human diseases (Pesu et al., 2008). The JAK family comprises of four non-receptor protein tyrosine kinases, namely JAK1, JAK2, JAK3, and TYK2 (Stark and Darnell, 2012). JAK1, JAK2, and TYK2 are expressed ubiquitously, while JAK3 is expressed mainly in hematopoietic cells.

Following the binding of a ligand to its receptor, receptor-associated JAKs are activated. Once activated by cytokines or growth factors through receptor-ligand interactions, JAKs phosphorylate the receptor and members of the STAT transcription factors. A number of STAT molecules, including STAT1, 3, 4, 5 and 6, have been identified (Murray, 2007) (Rawlings et al., 2004). STAT proteins once phosphorylated and activated by JAKs, dimerize and translocate to the nucleus where they modulate the expression of target genes (Vainchenker et al., 2011). Figure 1.13 shows the JAK-STAT signaling diagram (Liao et al., 2013).



Liao W, Lin JX, Leonard WJ Immunity Volume 38, Issue 1 2013 13 - 25

Figure 1.13 JAK-STAT signaling

Aberrant JAK-STAT signaling has been implicated in multiple human pathologies. The high incidence of acquired somatic activating mutations found in *JAK2* in MPNs (myeloproliferative neoplasms) is one example of the involvement of this pathway in disease. Mutations in the upstream thrombopoietin receptor (*MPLW515L*) and the loss of JAK regulation by LNK (LNK exon 2 mutations) have been associated with myelofibrosis (Vainchenker et al., 2011) (Pikman et al., 2006). Mutations in *JAK2*, mostly *JAK2 V617F*, that lead to constitutive activation of JAK2, have been found in the majority of patients with primary myelofibrosis (Kralovics et al., 2005) (Baxter et al., 2005) (Levine et al., 2005). Additional mutations in *JAK2* exon 12 have been identified in polycythemia vera and idiopathic erythrocytosis (Scott et al., 2007). Additionally, activated JAK-STAT has been implicated as a survival mechanism for human cancers (Hedvat et al., 2009). Taken together, multiple molecular mechanisms have been identified that can lead to aberrant activation of JAK-STAT signaling in human disease.

Given the roles of JAK-STAT activation in human cancers and the multiple ways in which the pathway can be dysregulated, it becomes important to identify patients with aberrantly activated JAK-STAT pathways that could benefit from JAK inhibitor therapy. The detection of JAK activation through the measurement of phospho-JAK or phospho-STATs in clinical samples is subject to many technical and logistical variables such as sample processing, epitope preservation and detection; therefore deriving a gene expression based signature indicative of STAT5 activation status could be of practical importance.

1.5 Thesis summary

This thesis involves the construction of a computational framework to determine the status of tumor suppressor genes and the application of this knowledge to the cell lines in the CCLE collection as outlined in the following steps:

1. Compilation of a list of tumor suppressor genes from the literature containing all previously validated tumor suppressor genes and also capturing some putative tumor suppressor genes. For genes on the list information was collected on known loss of function missense mutations.

2. Design and construction of a computational module to detect genetic alterations affecting both alleles. This module is based on analysis of data from SNP chips and data from exome sequencing.

3. Design and construction of a computational module to detect epigenetic alterations. Since epigenetic data is not wildly available, gene expression data is used as a proxy in most cases. Such substitution is not a perfect solution; however it is a reasonable practical approach. The expression data is used to identify samples with likely absent expression of gene in question.

4. Integration of the above modules in a flexible data analysis framework can allow users to determine tumor suppressor genes status and place them in one of the three categories described above in the introduction. Based on the category definitions and data used to place them in appropriate category the first category will contain genes that have high confidence of being completely disable in the cell line in question. The second category will contain genes which have medium confidence of being completely disabled in the cell line in question. The

third category will contain genes which have lower confidence of being completely disabled in the cell line in question.

5. The framework was used to analyze CCLE collection of cancer cell lines. The initial results of the analyses performed in this study have already been made available for scientific community through publication in a leading journal (Sonkin et al., 2013).

The construction of computational framework to determine gene sets activity across an extensive collection of gene sets and it application to the cell lines of the CCLE collection consisted of the following steps:

1. Determination of the appropriate normalization approach for expression of genes a across set of samples.

2. Determining an appropriate method for measuring gene sets activity scores using normalized data points from the first step.

3. Designing a computational approach to calculate statistical significance of gene sets activity scores.

4. Validation of gene sets activity scores on the set of gene sets with known behavior in a particular set of samples.

5. The framework was used to analyze gene sets activity across the CCLE collection of cancer cell lines using a large collection of annotated gene sets. Results were used as one of the inputs for predictive modeling of anticancer drug sensitivity (Barretina et al., 2012).

6. This underlying method was used in part to establish pSTAT5 mRNA expression signature in hematopoietic cancer cell lines and manuscript is under review (Sonkin et al., 2014).

Chapter Two: Tumor suppressor genes status in Cancer Cell Line Encyclopedia

This chapter presents details of systematic and comprehensive computational framework for the assessment of tumor suppressor genes status developed as part of thesis research.

Chapter 2.1 provides rational for selecting particular tumor suppressor genes for analysis and collection of associated annotation data. Chapter 2.2 describes the approach used to incorporate several orthogonal genomic data types, such as mutation data, copy number, LOH and expression in order to account for different mechanisms of tumor suppressor genes inactivation mechanisms. Chapter 2.3.1 presents major findings. Chapter 2.3.2 uses relationship between TP53 status and Nutlin-3 sensitivity in order to demonstrate that integrative method, which allows accounting for multiple mechanisms of tumor suppressor genes inactivation, provides a more accurate assessment of tumor suppressor genes status than can be inferred by expression, copy number, or mutation alone. Chapters 2.3.3 uses relationship between RB1 status and PD-0332991 sensitivity to reiterate conclusion from chapter 2.3.2. Chapter 2.4.1 highlights clinical relevance of TP53 status, chapter 2.4.2 uses BAP1 example to show how integrative tumor suppressor gene status can help to find new putative tumor suppressor genes and chapter 2.4.3 outlines potential future refinements.

2.1 Selecting tumor suppressor genes for analysis

The list of 82 well-known and putative tumor suppressors has been compiled based on a comprehensive literature review. Among them, 69 genes have mutation, copy number and expression data available and, therefore, were selected for the analysis in this thesis. List of these 69 genes is provided in Appendix A1.2.

The information was assembled from the literature on known loss of function missense mutations Table 2.1. Unfortunately, at this time the number of clearly validated loss of function missense mutations is small (only 38 entries covering 7 genes). It is likely that there are other *bona fide* losses of function missense mutations that have not been sufficiently validated or annotated. Creation of a comprehensive and reliable resource of clearly validated loss of function (as well as gain of function) missense mutations in cancer would be very useful for the field of oncology research. Hopefully, efforts like the MutaDATABASE (Bale et al., 2011) and ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) projects will help to establish such a resource. Also, as can be seen in the Table 2.1, some of the validated somatic loss of function mutations are found in dbSNP (Sherry et al., 2001), and are not explicitly noted to be somatic variants or pathogenic alleles. Therefore, in order to prevent the incorrect removal of somatic mutations, extra steps must be taken when using dbSNP as a filter to remove germ line SNPs from cancer samples sequencing.

Table 2.1 Known loss of function missense mutations.

Gene	ENTREZ ID	AA Change	dbSNP ID	Dominant Negative
CDKN2A	1029	H83Y		
CDKN2A	1029	D84Y	rs11552822	
CDKN2A	1029	D108Y		
CDKN2A	1029	P114L		
MLH1	4292	V384D		
PTEN	5728	R130G		
PTEN	5728	R130Q		
PTEN	5728	R173C		
PTEN	5728	R173H		
RB1	5925	C706F		
STK11	6794	D194N		
STK11	6794	D194V		
STK11	6794	E199K		
STK11	6794	P281L		
TP53	7157	V143A		N
TP53	7157	V157F		Y
TP53	7157	R158L		Y
TP53	7157	R158H		N
TP53	7157	R175H	rs28934578	Y
TP53	7157	Y220C		Y
TP53	7157	M237I		N
TP53	7157	G245S	rs28934575	Y
TP53	7157	R248Q	rs11540652	Y
TP53	7157	R248W		Y
TP53	7157	R249S		Y
TP53	7157	R273C		Y
TP53	7157	R273H	rs28934576	Y
TP53	7157	R273L		Y
TP53	7157	R280K		N
TP53	7157	R280S		N
TP53	7157	R280T		N
TP53	7157	R282G		N
TP53	7157	R282W		Y
VHL	7428	P81S	rs5030806	
VHL	7428	L85P	rs5030828	
VHL	7428	L89H	rs5030807	
VHL	7428	L158Q		
VHL	7428	R167W	rs5030820	

2.2 Methods

2.2.1 Overview

As described above mechanisms of inactivation of tumor suppressors can be divided into three major categories. Figure 2.1 illustrates each sub-category with a simplified diagram.

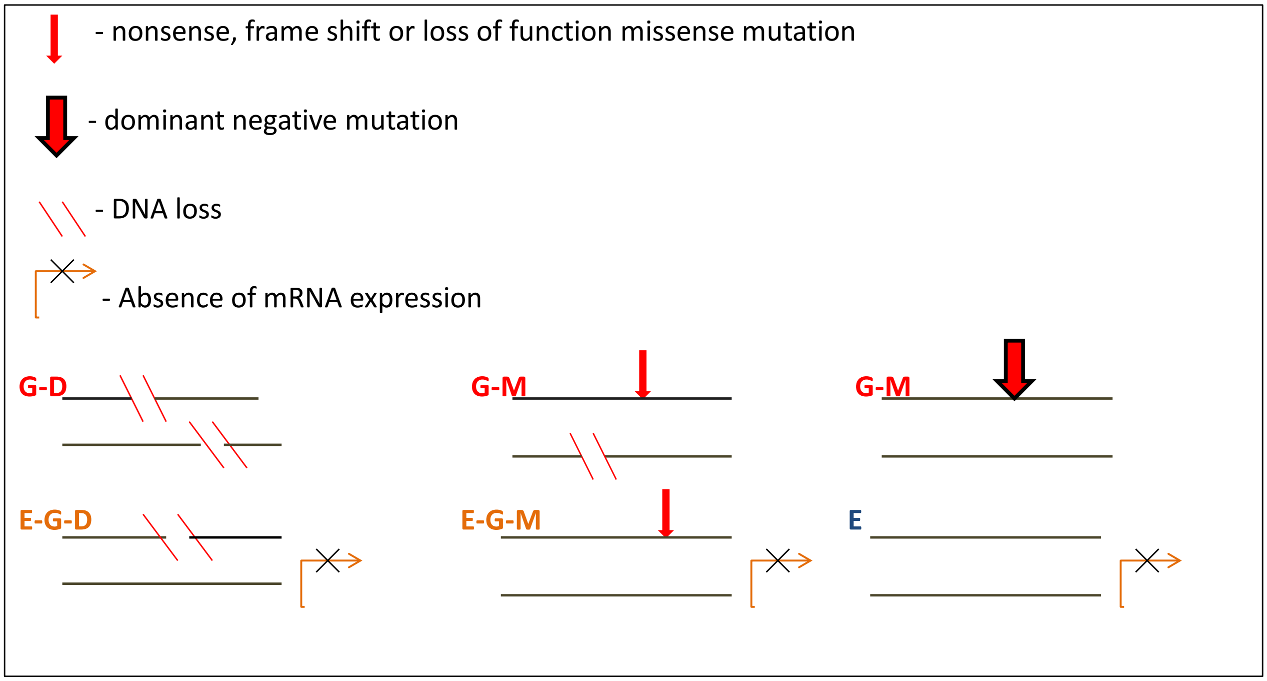

- nonsense, frame shift or loss of function missense mutation
- dominant negative mutation
- DNA loss
- Absence of mRNA expression
G-D
G-M
G-M
E-G-D
E-G-M
E

Figure 2.1 Tumor suppressor inactivation categories.
G - stands for genetic alteration, **D** - stands for deletion, **M** – stands for mutation
E - stands for absence of expression

2.2.2 Inactivation category by genetic mechanisms

The first inactivation category “G” is based completely on genetic mechanisms of inactivation of both alleles (Stanbridge, 1990) (Ponder, 2001) and, therefore, can be considered as the highest confidence category.

The genetic category can be subdivided further into 2 sub-categories:

1. The sub-category “G-M” is based on a homozygous nonsense, frame shift, loss of function missense mutation or heterozygous/homozygous dominant negative mutation.

2. The sub-category “G-D” is based on deletion of both alleles (bi-allelic loss).

One way for a gene to appear in the sub-category “G-M” is to have LOH status derived from Affymetrix SNP 6.0 data and a homozygous mutation deduced from the exome sequencing data. The exons of targeted genes were covered to average depth of 60-fold. A strict cut off of at least twenty reads for the mutant allele is applied to the exome data in order to decrease the possibility of failure to obtain sequence data for both alleles due to under sampling; no more than one read for the wild type allele is allowed. Any nonsense or frame shift mutation is considered to lead to the loss of function; only validated loss of function missense mutations from the Table 2.1 are used. Figure 2.1 illustrates a sub-category “G-M” with the most likely scenario of loss of one allele and inactivation of the other by mutation. However it is possible to have identical multiple copies of an allele inactivated by the mutation. It is useful to keep in mind that males have an automatic genetic “LOH” on X chromosome, and females have a mosaic allele specific expression pattern (due to random inactivation of one of two X chromosomes during early embryogenesis (Heard et al., 1997)).

Another “G-M” mechanism is dominant negative mutation. As can be seen from Table 2.1, only TP53 is considered to have dominant negative mutations in the list of 69 tumor suppressors that are examined in this paper. A cut-off value of at least 20 reads for the mutant allele in exome sequencing data is used to consider mutation as a trusted one. This stringent cut off is used to minimize false positive calls. OncoMap mutation calls are also used for this sub-category.

Exome sequencing data generated using short reads has a variable read depth with highest coverage approximately in the middle of capture probes and gradually decreasing coverage as the distance increases. Coverage drop at the TP53 R273C mutation (Figure 2.2) is example of such scenario and something that needs to be kept in mind while working with short reads exome sequencing data. In Figure 2.2 Broad institute Integrative Genomics Viewer (IGV) (Robinson et al., 2011) is used to show alignment of short reads to a 374 bp fragment of TP53 reference sequence in the SW-1710 cell line. On the far right side of the Figure 2 red double headed line points to the 14 reads with green bar representing nucleotide change causing R273C amino acid change (opposite end of the line points to the corresponding reference sequence). The top part of the figure shows a bar histogram representing the depth of coverage. The histogram shows maximum reads depth of 147 around 7,576,900 bp and coverage depth dropping below 20 around 7,577,100 bp. In the particular case of TP53

R273C mutation the data provided by OncoMap for that mutation helps to augment sequencing data.

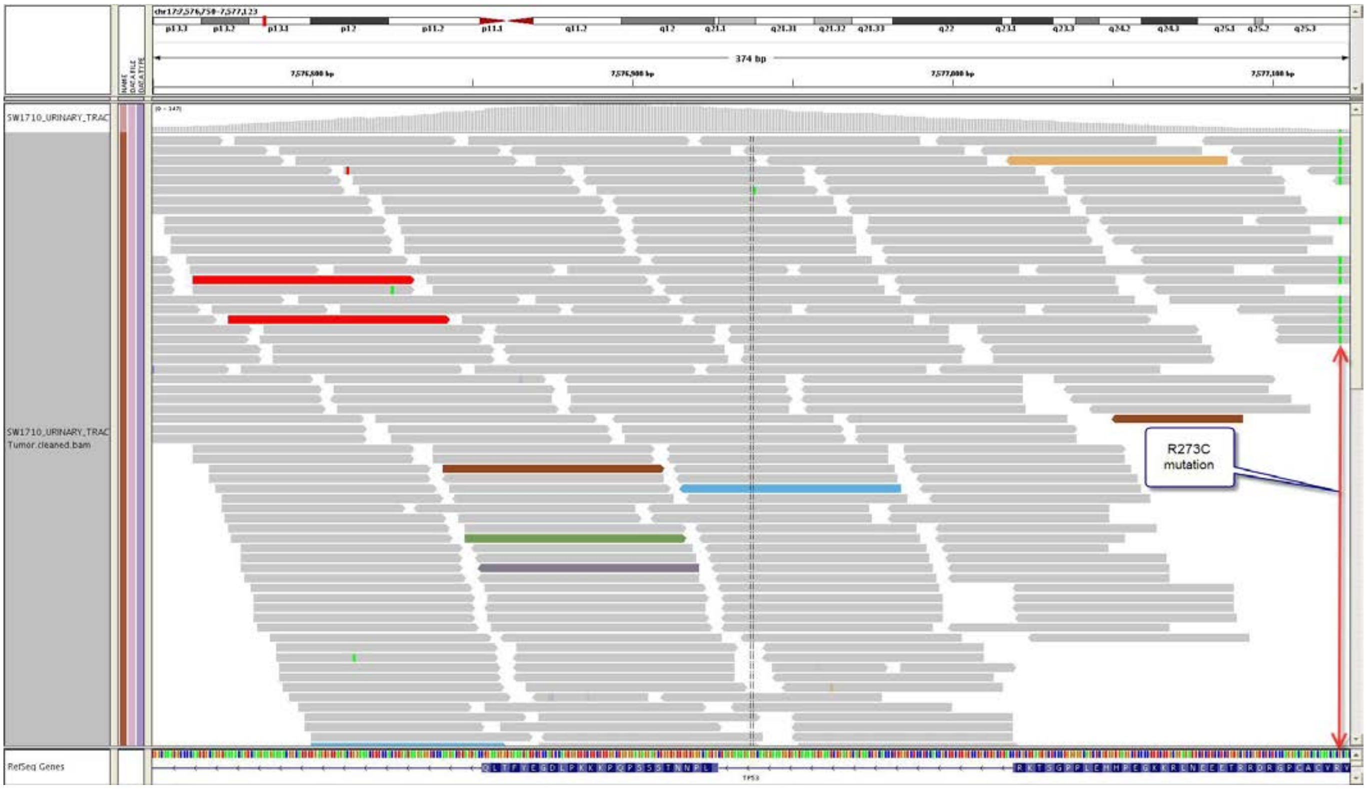


Figure 2.2 Coverage drop at the TP53 R273C mutation in SW1710 cell line.
IGV is used to show alignment of reads to 374 bp fragment of TP53 reference sequence in SW1710 cell line.
On the far right side of the figure red double headed line points to the 14 reads with green bar representing nucleotide change causing R273C amino acid change (opposite end of the line points to the corresponding reference sequence). Histogram shows maximum reads depth of 147 around 7,576,900 bp and coverage depth drop below 20 around 7,577,100 bp.

For a gene to be in the sub-category “G-D”, it must have a loss of both alleles based on the Affymetrix SNP 6.0 data. Theoretically, a CN ratio of 1 corresponds to a diploid genome with 2 copies of the gene, CN ratio of 1.5 corresponds to 3 copies of the gene, CN ratio of 0.5 corresponds to 1 copy of the gene, etc. Since there is data compression and increase in noise at low CN ratio numbers, CN ratio below 0.25 is used as an indicator of loss of both alleles.

2.2.3 Inactivation category by Genetic mechanisms and loss of the expression

The second category “E-G” includes inactivation of one allele by a genetic mechanism and loss of the expression of the second allele. The loss of the expression could be due to multiple reasons, such as loss of upstream signaling, mutations in promoter and enhancer regions and any one of several possible epigenetic mechanisms, such as promoter methylation and possible histone modifications. The epigenetic mechanism of inactivation of tumor suppressor genes is considered to be of fundamental importance in tumorigenesis (Jones and Baylin, 2002). Since epigenetic data is not available at this point for the majority of the CCLE cell lines, gene expression data is used as a proxy for the epigenetic mechanism; this substitution is not perfect, however it provides a reasonable practical approach. Expression data is used to identify cell lines with likely absent expression of gene in question. Affymetrix U133Plus2 arrays have been used in CCLE to generate mRNA expression profiles. MAS5 algorithm with Target Signal Intensity set to 150 (Hubbell et al., 2002) is used to generate expression values. A given gene is considered as not expressed in the cell line if its expression is below 32, while both mean and median expression of this gene across all cell lines are above 100 indicating presence of dynamic range of expression. In general, expression below 32 in most cases would indicate the absence of the corresponding protein. For calculation of mean and median gene expression values, cell lines with CN ratio below 0.25 were discarded, in order to decrease artificial under-estimation of expression distributions of cell lines with remaining functional DNA.

An effort was made to use gene expression distribution between cell lines with CN ratio < 0.25 and cell lines with CN ratio > 0.6 in order to improve gene expression cut off. However, systematic improvement in threshold behavior has not been achieved. Attempts examining gene expression in particular lineages also did not result in systematic improvements, at least in part due to decrease in sample sizes. Also, as expected (Choe et al., 2005) (Pepper et al., 2007), the same overall results were obtained using RMA (Irizarry et al., 2003) instead of MAS5.

The second category can be further divided into two sub-categories:

1. The sub category “E-G-D” is characterized by deletion of one allele and absence of gene expression.

2. The sub category “E-G-M” is characterized by nonsense, frame shift or loss of function missense mutation on one allele and absence of gene expression.

Exome sequencing data and OncoMap mutation calls are used for this sub category. Since the second category, in general, requires absence of mRNA expression, sub category “E-G-M” will mostly cover the scenarios when the mutation leads to mRNA decay. Therefore, in the majority of cases mRNA expression from single allele with a loss of function mis-sense mutation will not qualify for sub category “E-G-M”.

Utilization of RNA-seq data (Mortazavi et al., 2008) would allow determination of allele-specific expression and, therefore, would help to better cover sub-category “E-G-M”. In the next phase of the CCLE project, RNA-seq data will be generated for the majority of cell lines.

2.2.4 Inactivation category by loss of the expression

The third category “E” is based on loss of the expression of both alleles. As in the “E-G” category, epigenetic mechanisms are likely playing an important role in the loss of the expression. Category “E-LOH” denotes LOH in addition to absence of mRNA expression.

In order to help identify cell lines with functional wild type tumor suppressors, wild type category was established. The wild type category is based on absence of non-synonymous mutations, splice sites mutations, CN loss (CN ratio above 0.9) or LOH (based on CN data).

The wild type category could be further divided into two sub-categories:

1. The sub category “WT-E” has mRNA expression of at least 300.
2. The sub category “WT” has mRNA expression above 32 and below 300.

Generally, mRNA expression above 300 indicates the presence of corresponding protein. Therefore, this cut-off is used for mRNA expression for the “WT-E” sub-category.

Finally, a catch-all category “0” is defined for cases which do not qualify any of the above categories. Category “0” has two additional sub-categories:

1. “0-D” is characterized by deletion of one allele.

2. “0-M” is characterized by heterozygous nonsense, frame shift or loss of the function missense mutation.

Figure 2.3 illustrates category assignment by means of simplified flow chart.

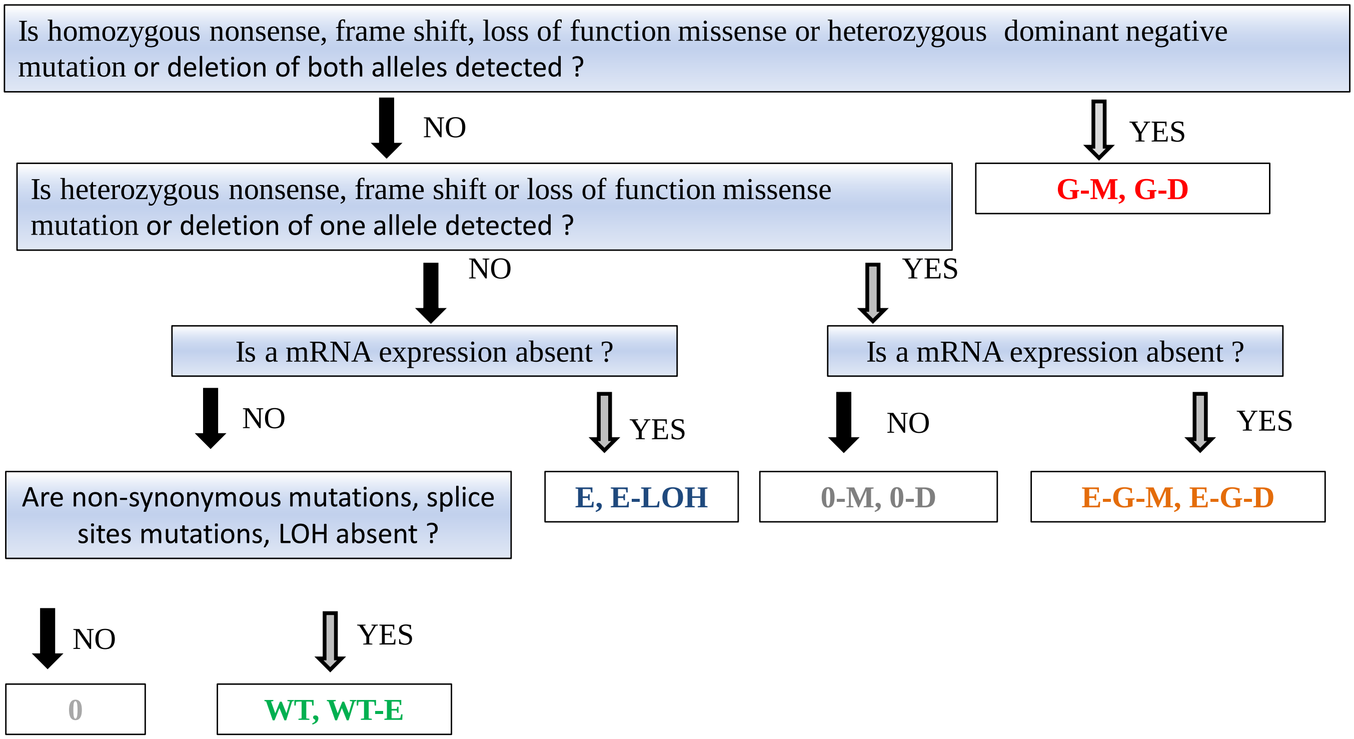

Is homozygous nonsense, frame shift, loss of function missense or heterozygous dominant negative mutation or deletion of both alleles detected ?
NO
YES
G-M, G-D
Is heterozygous nonsense, frame shift or loss of function missense mutation or deletion of one allele detected ?
NO
YES
Is a mRNA expression absent ?
Is a mRNA expression absent ?
NO
YES
NO
YES
Are non-synonymous mutations, splice sites mutations, LOH absent ?
E, E-LOH
0-M, 0-D
E-G-M, E-G-D
NO
YES
0
WT, WT-E

Figure 2.3 Simplified flow chart of category assignments.

2.3 Results

2.3.1 Overview

For 69 tumor suppressor genes which have all data types available a systematic and comprehensive matrix of tumor suppressor status across 799 CCLE cell lines using categories described in the Methods section have been generated. A web-interface for selecting cell lines with desirable status of tumor suppressor(s) is available at http://cancer.tools.glacombio.net. Appendix A1.3 lists summaries of inactivation categories counts for all examined genes and Table 2.2 shows fragment of Supplemental Excel file from (Sonkin et al., 2013).

Table 2.2 Status of 69 tumor suppressors in 799 cell lines.

Fragment of Supplemental Excel file from (Sonkin et al., 2013).

Name	APC	ATM	CDKN2A	CDKN2B	LATS2	MLH1	MSH2	NF2	PTEN	RB1	STK11	TP53	TSC1	VHL
697	0	WT-E	E-LOH	0	WT	WT-E	WT-E	0	WT-E	WT-E	WT	WT-E	WT	WT
BDCM	WT-E	0	WT	WT	WT	WT-E	WT-E	WT	WT-E	WT-E	WT	WT-E	WT	WT
BT-549	WT-E	WT	WT-E	WT	0	0	0	0	0	E-LOH	0	0	0	WT
GDM-1	WT-E	0	E	WT	WT-E	WT-E	WT-E	WT	WT-E	0	WT	WT-E	WT	WT
HPB-ALL	0	0	WT	WT	WT	G-M	0-D	WT-E	WT-E	WT-E	WT	E-G-D	WT	0-D
KE-37	WT-E	0	G-D	0	WT	WT-E	WT-E	WT-E	E-G-D	0	WT	WT	WT	WT
KOPN-8	0	WT-E	WT	WT	WT-E	0	WT-E	WT-E	WT-E	WT-E	WT	G-M	WT	WT-E
Loucy	0-D	0	0-D	0-D	E	WT-E	WT-E	WT	0	WT-E	WT	0	WT	WT
MHH-CALL-2	0	0	G-D	G-D	0	0	0	0	0	0	0	0	0	0
MHH-CALL-3	WT-E	WT-E	G-D	0-D	WT	WT-E	WT-E	WT-E	WT-E	WT-E	0-D	WT-E	WT-E	WT
MHH-CALL-4	WT-E	0	G-D	G-D	WT-E	WT-E	WT-E	0	WT-E	WT-E	WT	WT-E	WT	WT

Appendix A2 contains a CD with a CCLE_TS-Genes-Status.xls file which has tumor suppressor status for 69 genes across all 799 CCLE cell lines. Results presented in Appendix A1.3 clearly shows that a systematic and comprehensive framework to access tumor suppressor genes status is indeed able to capture multiple mechanisms of inactivation of tumor suppressor genes and therefore improves characterization of alterations in cancer cell lines. As one would expect, genes CDKN2A, TP53, RB1, PTEN and APC are among most frequently disabled genes in cancer cell lines as clearly demonstrated by Appendix A1.3. Not surprisingly, TP53 was inactivated in about 50 percent of cases by dominant negative mutations (Petitjean et al., 2007). CDKN2A appears to be the most frequently disabled gene in CCLE. In about 76% of cases CDKN2A is likely to be inactivated by DNA deletion of both alleles. While this is a larger proportion than may be expected, the effect may be partially explained by the small genomic size of the CDKN2A locus (only 25 kbp). Publications by (Kitagawa et al., 2002) and (Sasaki et al., 2003) also reported high frequency of CDKN2A inactivation by DNA deletion of both alleles. Also recent provisional TCGA data indicate that a high percentage of primary tumor samples have CDKN2A inactivated by DNA deletion of both alleles (http://www.cbioportal.org). This evidence seems to indicate that such an inactivation mechanism is not just specific to cancer cell lines, but may be commonly found in primary tumor samples. Figure 2.4 shows the relationship between CDKN2A CN ratios, mutation status and mRNA expression across 799 CCLE cancer cell lines. As can be seen from Figure 2.4 the frequency of CDKN2A inactivations by means of loss of mRNA expression without any other genetic alterations is not high. CDKN2B is disabled almost as often as CDKN2A, however CDKN2B is located just 6 kbp from CDKN2A locus and therefore the loci may be physically deleted together. Figure 2.5 shows the relationship between CDKN2A and CDKN2B CN ratios across 799 CCLE cancer cell lines.

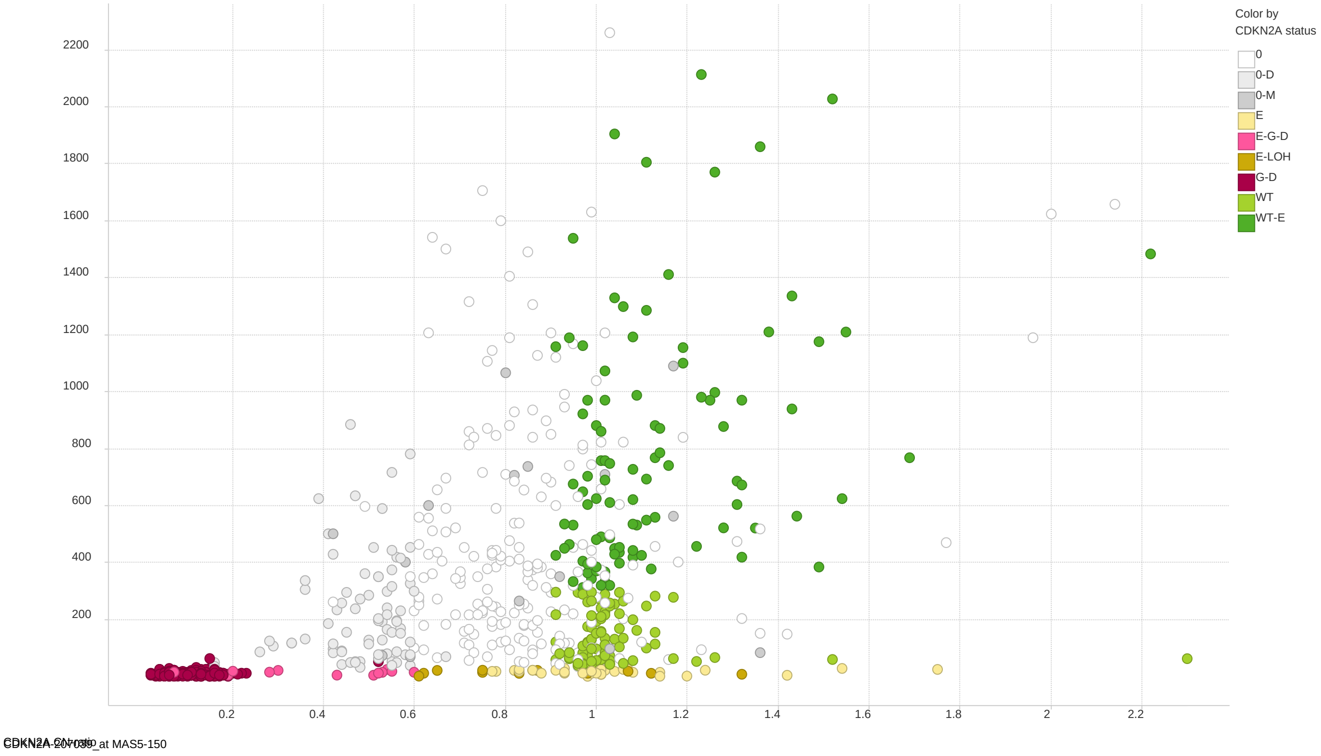


Figure 2.4 CDKN2A CN ratios, status and mRNA expression

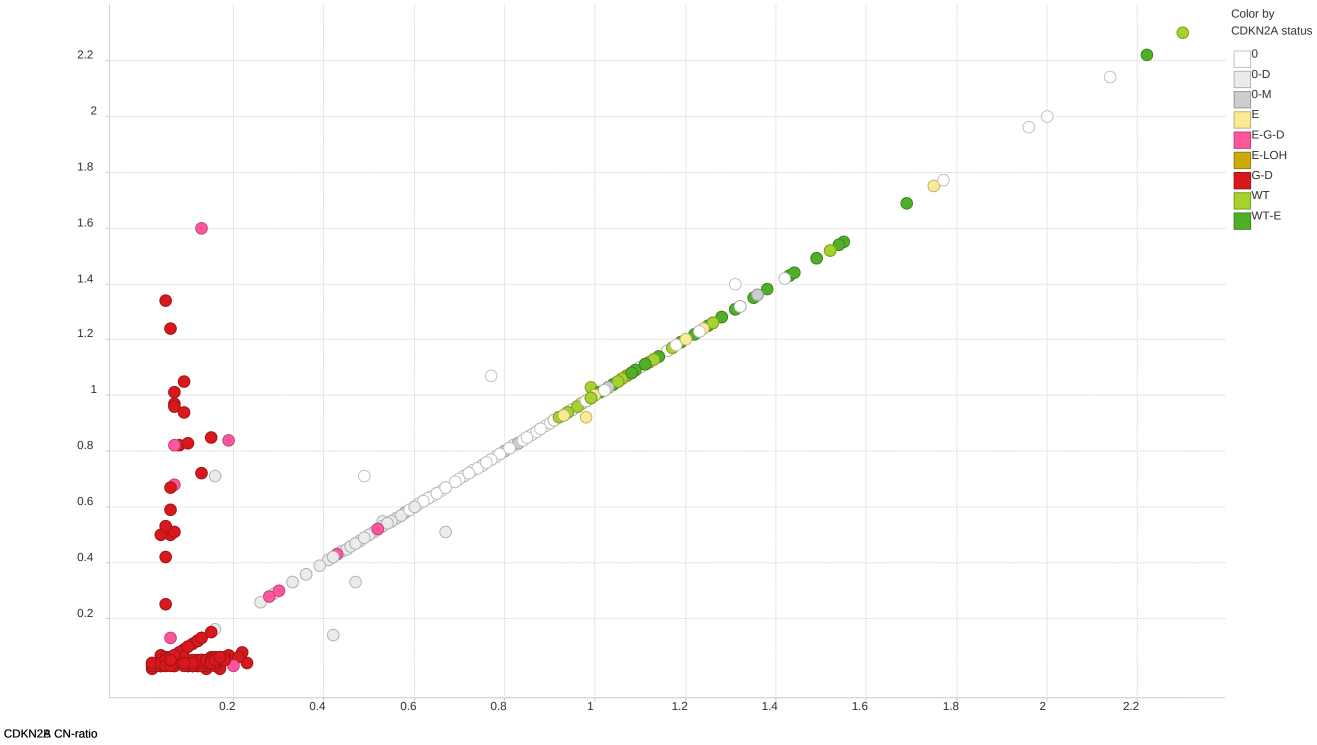


Figure 2.5 CDKN2A, CDKN2B CN ratios and CDKN2A status

2.3.2 TP53 status and Nutlin-3 sensitivity

Nutlin-3 is an inhibitor of MDM2-driven TP53 protein degradation (Vassilev, 2004) (Kubbutat et al., 1997). Mechanistically, only cell lines with wild type TP53 can be potentially sensitive to this inhibitor as confirmed in part by sensitivity of wild type MEFs cells and by the loss of sensitivity in the TP53 knock out MEFs (Efeyan et al., 2007). This knowledge can be used to assess an integrative approach to determining tumor suppressor status. Figure 2.6 shows the relationship between TP53 status and sensitivity to Nutlin-3. Cell lines with IC50 below 4.5 µM are considered to be sensitive to Nutlin-3 and cell lines with IC50 above 6.5 µM are considered to be insensitive to Nutlin-3. (The IC50 is concentration at which the drug response reached an absolute inhibition of 50%). As illustrated by Figure 2.6, all cell lines with TP53 inactivated by any mechanism are insensitive to Nutlin-3. Table 2.3 shows that each inactivation category is statistically significantly enriched for insensitive cell lines in comparison to wild type ones, by the Fisher exact test.

Table 2.3 also illustrates that combining multiple inactivation categories leads to more significant statistical results. In cell lines with inactivated TP53 one would expect to see a drop in TP53-driven signaling. The KEGG (Kanehisa et al., 2012) and BIOCARTA (http://www.biocarta.com) representations of the TP53 signaling pathway are well established references. Pathway activity scores were calculated for 4,162 MSigDB (Liberzon et al., 2011) gene sets covering multiple gene sets sources including KEGG and BIOCARTA using the approach described in Chapter 3. Correlation coefficients were calculated between each TP53 inactivation category / wild type category and TP53 pathway activity scores based on KEGG and BIOCARTA gene sets.

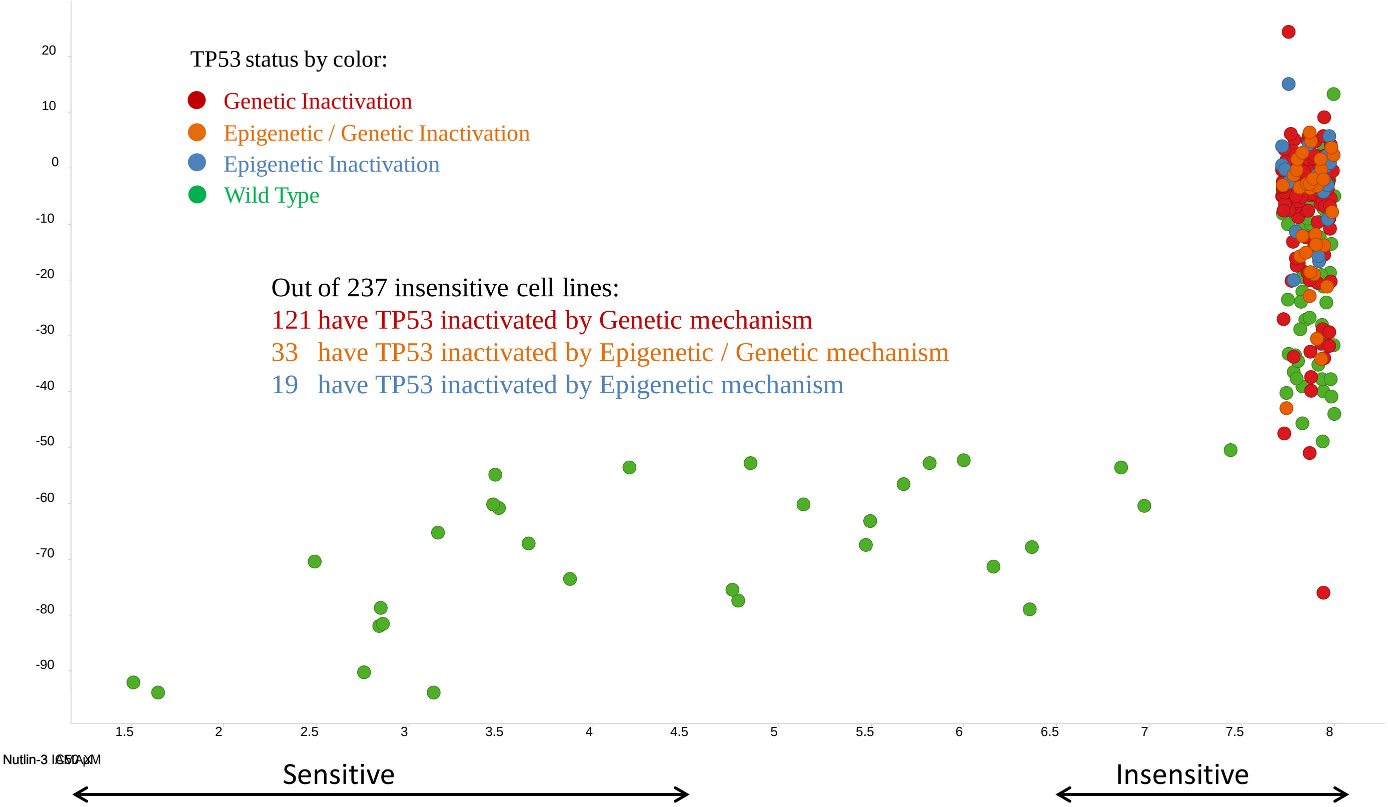


Figure 2.6 Nutlin-3 sensitivity across 491 CCLE cell lines in relation to TP53 status
Each dot represents cell line with TP53 inactivation status marked by color.
All cell lines with TP53 inactivated by any mechanism are insensitive to Nutlin-3.
TP53 Wild Type status covers: WT-E and WT categories. TP53 Genetic Inactivation status covers: G-M and G-D categories. TP53 Epigenetic Inactivation status covers E-LOH category.
TP53 Epigenetic / Genetic Inactivation status covers: E-G-M and E-G-D categories.

Table 2.3 TP53 inactivation categories and Nutlin-3 insensitivity.

TP53 inactivation	Fisher exact p-value for Nutlin-3 insensitivity
Genetic Inactivation	3.69E-07
Epigenetic / Genetic Inactivation	3.43E-03
Epigenetic Inactivation	2.95E-02
Epigenetic / Genetic Inactivation and Epigenetic Inactivation	2.82E-04
Epigenetic / Genetic Inactivation, Epigenetic Inactivation and Genetic Inactivation	1.03E-08

p-values indicate how likely just by chance the observed enrichment in insensitive cell lines for each of inactivation groups vs. wild type cell lines.
Genetic Inactivation covers: G-M and G-D categories.
Epigenetic / Genetic Inactivation covers: E-G-D category.
Epigenetic Inactivation covers E-LOH category.

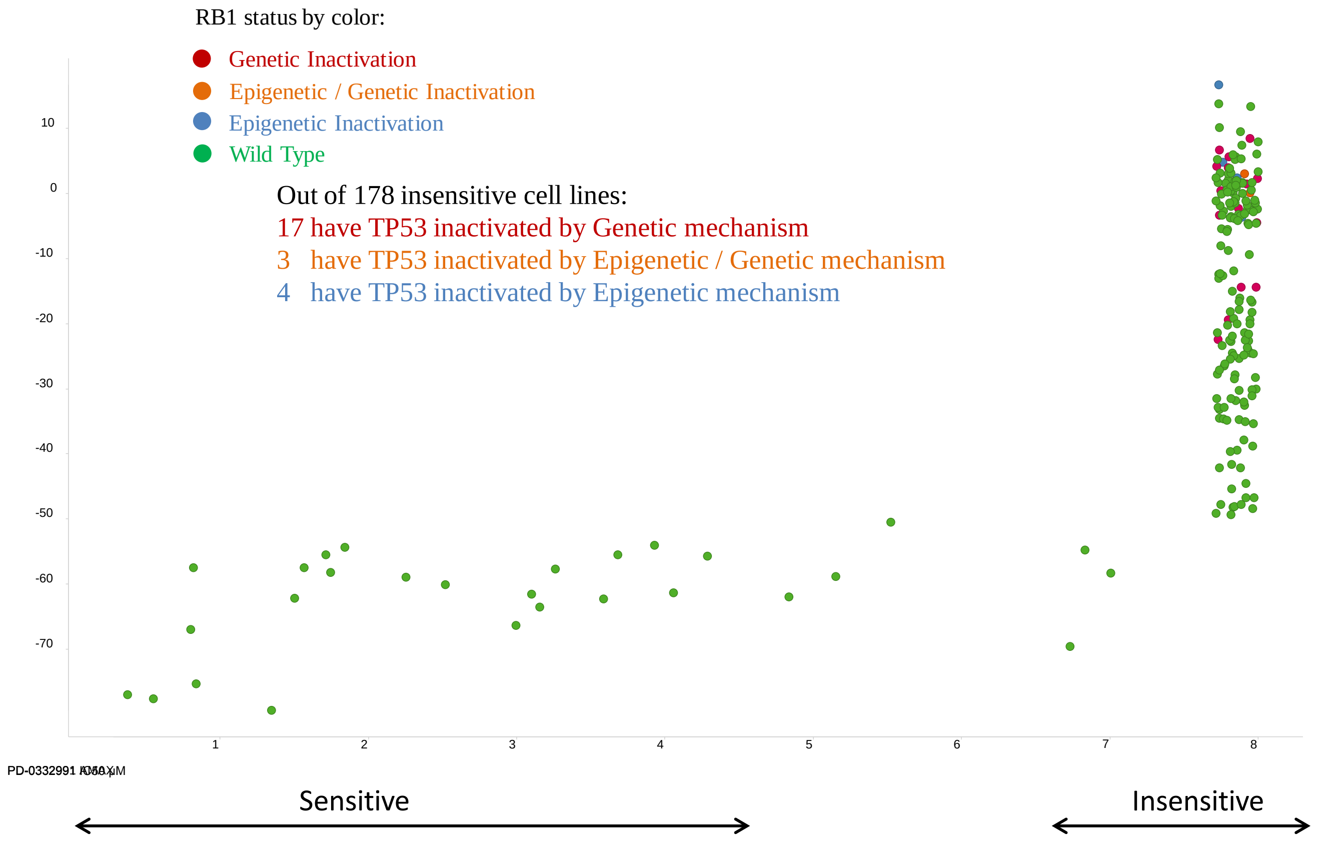
Table 2.4 summarizes results of correlation calculations, as anticipated negative correlations are observed for each inactivation category. Out of 4,162 correlation coefficients for each inactivation category, the correlation coefficients for BIOCARTA TP53 pathway have the most negative values in 4 out of 5 cases. The statistical significance of the distribution of TP53 pathway activity scores between cell lines from particular inactivation category and wild type cell lines is confirmed by t-test. As indicated by above observations, the consideration of multiple mechanisms of inactivation provides us with a more complete and informative landscape of TP53 inactivation and better genomic characterizations of cell lines models.

Table 2.4 TP53 inactivation categories and TP53 Signaling Pathways.

TP53 inactivation	**Correlation coefficient and rank for KEGG TP53 Signaling Pathway**	**t-test p-value for KEGG TP53 Signaling Pathway scores**	**Correlation coefficient and rank for BIOCARTA TP53 Signaling Pathway**	**t-test p-value for BIOCARTA TP53 Signaling Pathway scores**
Genetic Inactivation	-0.35 (2)	1.15E-10	-0.45 (1)	2.41E-17
Epigenetic / Genetic Inactivation	-0.35 (4)	4.40E-07	-0.45 (2)	1.42E-11
Epigenetic Inactivation	-0.39 (2)	2.67E-08	-0.51 (1)	2.47E-14
Epigenetic / Genetic Inactivation and Epigenetic Inactivation	-0.44 (3)	1.77E-12	-0.57 (1)	3.18E-21
Epigenetic / Genetic Inactivation, Epigenetic Inactivation and Genetic Inactivation	-0.39 (2)	3.95E-16	-0.51 (1)	2.54E-27

2.3.3 RB1 status and PD-0332991 sensitivity

PD-0332991 is a CDK4/6 inhibitor and mechanistically only cell lines with wild type RB1 can be potentially sensitive to this inhibitor (Finn et al., 2009). Figure 2.7 shows the relationship between RB1 status and sensitivity to PD-0332991. Cell lines with IC50 below 4.5 µM are considered to be sensitive to PD-0332991 while cell lines with IC50 above 6.5 µM are considered to be insensitive to PD-0332991.



Out of 178 insensitive cell lines:
17 have TP53 inactivated by Genetic mechanism
3 have TP53 inactivated by Epigenetic / Genetic mechanism
4 have TP53 inactivated by Epigenetic mechanism

Figure 2.7 PD-0332991 sensitivity across 203 CCLE cell lines in relation to RB1 status.
Each dot represents cell line with RB1 inactivation status marked by color.
All cell lines with RB1 inactivated by any inactivation mechanism are insensitive to PD-0332991.
RB1 Wild Type status covers: WT-E and WT categories. RB1 Genetic Inactivation status covers: G-M and G-D categories.
RB1 Epigenetic Inactivation status covers E-LOH category. RB1 Epigenetic / Genetic Inactivation status covers: E-G-D category.

As illustrated by Figure 2.7 all cell lines with RB1 inactivated by any mechanism are insensitive to PD-0332991. Table 2.5 shows that each inactivation category by itself is not statistically significantly enriched for insensitive cell lines in comparison to wild type ones, however statistical significance is reached by combining multiple inactivation categories.

Table 2.5 RB1 inactivation categories and PD-0332991 insensitivity.

RB1 inactivation	**Fisher exact p-value for PD-0332991 insensitivity**
Genetic Inactivation	1.16E-01
Epigenetic / Genetic Inactivation	6.73E-01
Epigenetic Inactivation	5.91E-01
Epigenetic / Genetic Inactivation and Epigenetic Inactivation	4.01E-01
Epigenetic / Genetic Inactivation, Epigenetic Inactivation and Genetic Inactivation	5.06E-02

p-values indicate how likely just by chance the observed enrichment in insensitive cell lines for each of inactivation groups vs. wild type cell lines.
Genetic Inactivation covers: G-M and G-D categories.
Epigenetic / Genetic Inactivation covers: E-G-D category.
Epigenetic Inactivation covers E-LOH category.

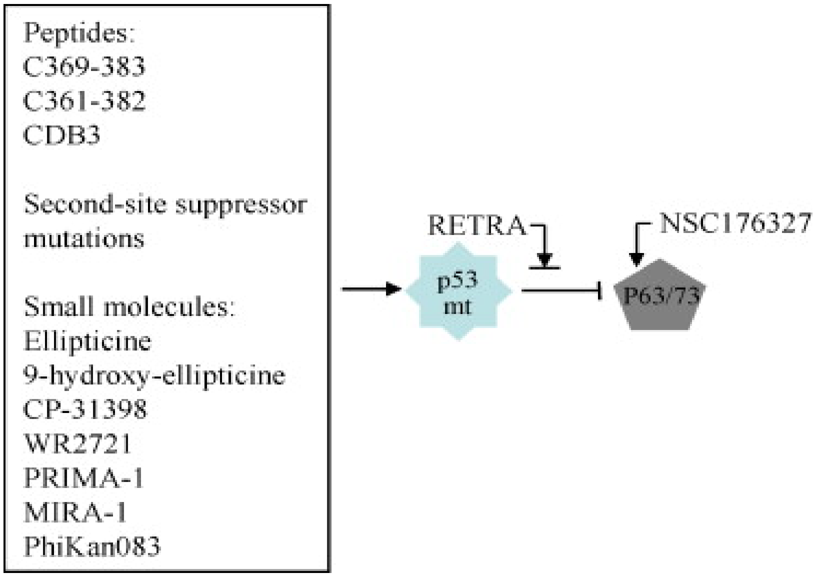
2.4 Discussion

The examples in Chapters 2.3.2 and 2.3.3 for TP53 and RB1 clearly demonstrate that accounting for multiple mechanisms of tumor suppressor genes inactivation leads to a much more accurate determination of tumor suppressor functional status. Such enhanced accuracy could be useful component in efforts to improve preclinical stratification of anticancer therapeutics. A summary of these results has been published in a leading journal (Sonkin et al., 2013).

The approach presented here and its application to CCLE will clearly improve the characterization of the status of tumor suppressor genes in cancer cell line models. It is important to reiterate that the status of tumor suppressor gene(s) can play a critical role in selecting appropriate therapeutic strategy for the patient in clinical practice. In order to highlight this point it is useful to take more detailed look at strategies to target TP53 in cancer.

2.4.1 TP53 clinical relevance

In about 50% of cancer cases TP53 is directly inactivated by number of different mechanisms and in approximately other 50% of cancer cases wild type TP53 protein function is at least partially inhibited or TP53 signaling pathway is down regulated (Brown et al., 2009). Work in multiple cancer models demonstrated that absence of TP53 function is important in tumor maintenance and therefore restoring TP53 function is promising therapeutic approach (Martins et al., 2006) (Ventura et al., 2007) (Xue et al., 2007). Globally there are about 13 million patients diagnosed with cancer every year (Jemal et al., 2011), therefore about 6.5 million patients each year are diagnosed with tumor(s) in which TP53 is inactivated and approximately other 6.5 million patients each year are diagnosed with tumor(s) with wild type TP53 protein. Therefore strategies to target TP53 in cancer are split in two major groups. One set of strategies is for targeting cancer with no wild type TP53 protein and the second set of strategies is for targeting cancer with wild type TP53 protein. Figure 2.8 summarizes approaches to target tumors with mutant TP53 protein and Figure 2.9 summarizes approaches to target tumors with wild type TP53 protein (Chen et al., 2010).



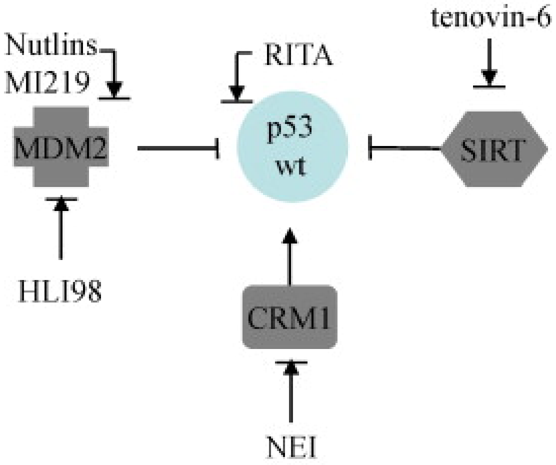
Fang Chen , Wenge Wang , Wafik S. El-Deiry Biochemical Pharmacology Volume 80, Issue 5 2010 724 - 730

Figure 2.8 Strategies in reactivation of mutant TP53.

NSC176327 induces p73 expression and activates TP53-like activity in a TP53 independent manner. This approach is especially useful in tumors with TP53 deletion.

RETRA relieves the negative effect of mt TP53 on p73 and induces p73 expression. Treatment of RETRA exhibits TP53-like activity.

All molecules/approaches in the box rescue mt TP53 function by assisting mt TP53 refolding or driving the protein folding equilibrium from denatured/mt conformation more toward the native functional/wt conformation. Restoration of native conformation to mt TP53 endows its wild-type activity.



Fang Chen , Wenge Wang , Wafik S. El-Deiry Biochemical Pharmacology Volume 80, Issue 5 2010 724 - 730

Figure 2.9 Strategies in wild-type TP53 activation.

Tenovin-6 inhibits protein-deacetylase activity of SirT1 and SirT2. Acetylation results in the stabilization of TP53 and interferes MDM2-mediated degradation.

NEI inhibits the nuclear export protein CRM1 which increases nuclear TP53 level indirectly.

HLI98 blocks the ubiquitin ligase activity of HDM2. It prevents TP53 degradation and elevates TP53 level indirectly.

RITA binds to TP53 and interferes with MDM2 and TP53 interaction which activates TP53 function.

Both Nutlins and MI219 interact with MDM2 and block TP53 interaction. Consequently, they activate TP53.

In addition to the approaches outlined in Figure 2.8 and Figure 2.9 the TP53 gene therapy (Huang et al., 2009) (Yang et al., 2010) and synthetic lethal approaches (Brown et al., 2009) are the other major areas of the research.

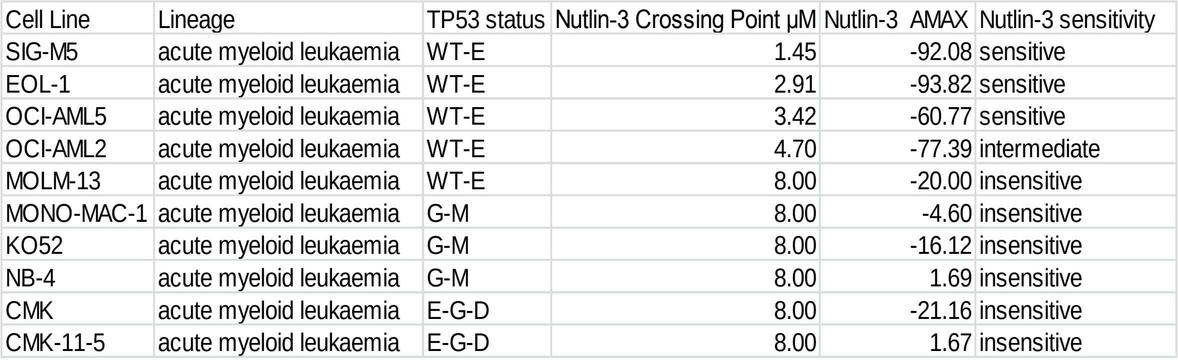
MDM2 TP53 interaction inhibitors are the most clinically advanced therapeutics for trying to improve treatment for patients with wild type TP53 protein. Table 2.6 lists MDM2 TP53 interaction inhibitors which are currently in early clinical trials. (Source http://clinicaltrials.gov)

Table 2.6 MDM2 TP53 interaction inhibitors in clinical trials

Compound Name	Sponsoring Organization	Phase
RO5045337	Hoffmann-La Roche	Phase 1B Dose Escalation Study
RO5503781	Hoffmann-La Roche	Phase 1 Dose Escalation Study
MK-8242	Merck	Phase 1 Dose Escalation Study
SAR405838	Sanofi	Phase 1 Dose Escalation Study
AMG 232	Amgen	Phase 1 Dose Escalation Study
CGM097	Novartis Pharmaceuticals	Phase 1 Dose Escalation Study

Table 2.7 shows sensitivity to MDM2 TP53 interaction inhibitor Nutlin-3 in CCLE Acute myeloid leukemia (AML) cell lines.

Table 2.7 Sensitivity to Nutlin-3 in CCLE AML cell lines.



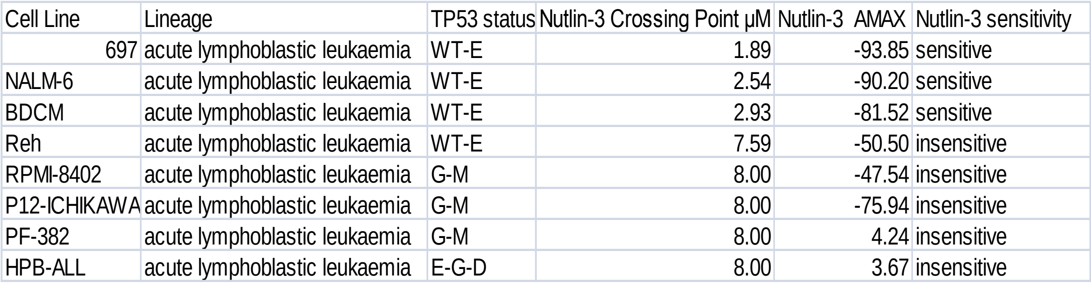
Cell Line	Lineage	TP53 status	Nutlin-3 Crossing Point µM	Nutlin-3 AMAX	Nutlin-3 sensitivity
SIG-M5	acute myeloid leukaemia	WT-E	1.45	-92.08	sensitive
EOL-1	acute myeloid leukaemia	WT-E	2.91	-93.82	sensitive
OCI-AML5	acute myeloid leukaemia	WT-E	3.42	-60.77	sensitive
OCI-AML2	acute myeloid leukaemia	WT-E	4.70	-77.39	intermediate
MOLM-13	acute myeloid leukaemia	WT-E	8.00	-20.00	insensitive
MONO-MAC-1	acute myeloid leukaemia	G-M	8.00	-4.60	insensitive
KO52	acute myeloid leukaemia	G-M	8.00	-16.12	insensitive
NB-4	acute myeloid leukaemia	G-M	8.00	1.69	insensitive
CMK	acute myeloid leukaemia	E-G-D	8.00	-21.16	insensitive
CMK-11-5	acute myeloid leukaemia	E-G-D	8.00	1.67	insensitive

As was highlighted in section 2.3, independently of the mechanism of inactivation of TP53 all cell lines with disabled TP53 are insensitive to Nutlin-3. In the panel of CCLE AML cell lines approximately half of TP53 wild type cell lines are sensitive to Nutlin-3. In the TCGA study across 200 primary AML samples TP53 was inactivated in about 9% of cases (Cancer

Genome Atlas Research Network, 2013). Currently the first-line treatment of most AML subtypes consists primarily of chemotherapy. Worldwide there are estimated 200,000 new cases of AML each year (Jemal et al., 2011), indicating that MDM2 TP53 interaction inhibitors could potentially improve treatment for approximately 90,000 new AML patients each year.

Table 2.8 shows sensitivity to MDM2 TP53 interaction inhibitor Nutlin-3 in CCLE Acute lymphoblastic leukemia (ALL) cell lines.

Table 2.8 Sensitivity to Nutlin-3 in CCLE ALL cell lines.



Cell Line	Lineage	TP53 status	Nutlin-3 Crossing Point µM	Nutlin-3 AMAX	Nutlin-3 sensitivity
697	acute lymphoblastic leukaemia	WT-E	1.89	-93.85	sensitive
NALM-6	acute lymphoblastic leukaemia	WT-E	2.54	-90.20	sensitive
BDCM	acute lymphoblastic leukaemia	WT-E	2.93	-81.52	sensitive
Reh	acute lymphoblastic leukaemia	WT-E	7.59	-50.50	insensitive
RPMI-8402	acute lymphoblastic leukaemia	G-M	8.00	-47.54	insensitive
P12-ICHIKAWA	acute lymphoblastic leukaemia	G-M	8.00	-75.94	insensitive
PF-382	acute lymphoblastic leukaemia	G-M	8.00	4.24	insensitive
HPB-ALL	acute lymphoblastic leukaemia	E-G-D	8.00	3.67	insensitive

In the panel of CCLE ALL cell lines approximately 75% of TP53 wild type cell lines are sensitive to Nutlin-3. In primary ALL samples TP53 is inactivated in about 30% of cases and inactivation mechanisms are split approximately equally between point mutations, DNA loss and loss of expression (Agirre et al., 2003). Currently the first-line treatment of most ALL subtypes consists primarily of chemotherapy. Worldwide there are estimated 120,000 new cases of ALL each year (Jemal et al., 2011), indicating that MDM2 TP53 interaction inhibitors could potentially improve treatment for approximately 60,000 new ALL patients each year.

Table 2.9 lists clinical trials indications for MDM2 TP53 interaction inhibitors. It is evident from this table that there is a wide variety of indications in which initial wave of clinical trials is performed.

Table 2.9 Clinical trials indications for MDM2 TP53 interaction inhibitors.

Indication
Acute Lymphoblastic Leukemia
Acute Myeloid Leukemia
Chronic Lymphocytic Leukemia
Chronic Myeloid Leukemia
Hodgkin Lymphoma
Liposarcoma
Melanoma
Non-Hodgkin Lymphoma
Prostate
Soft Tissue Sarcomas
Solid Tumors

Different indications have varying frequencies of TP53 wild type tumors and also varying sensitivity to MDM2 TP53 interaction inhibitors. Figure 2.10 shows TP53 status across CCLE tumor sites. As mention before 6.5 million new patients each year are estimated to have tumors with wild type TP53. Therefore there is a potential for MDM2 TP53 interaction inhibitors to improve treatment for hundreds of thousands new patients each year.

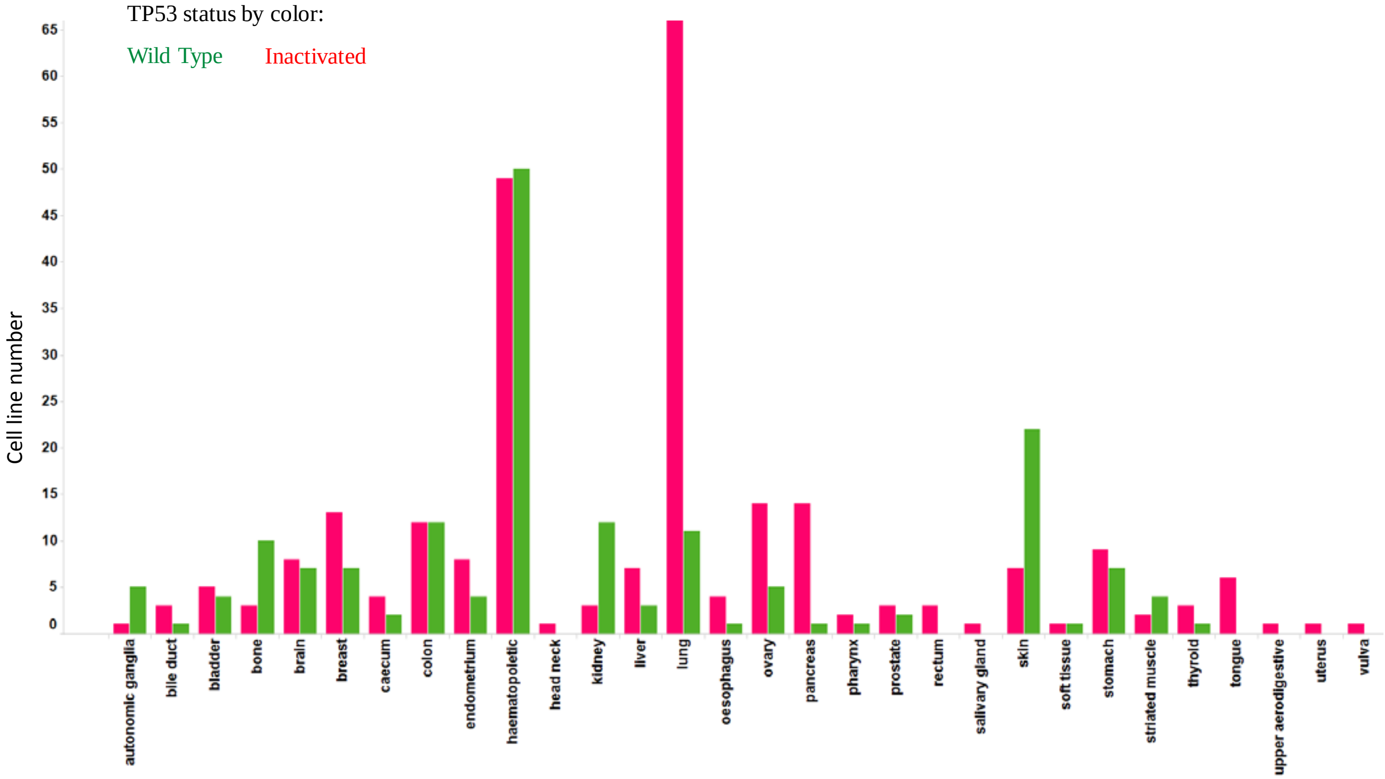


Figure 2.10 TP53 status across CCLE tumor sites.

TP53 Wild Type status covers: WT-E and WT categories.

TP53 Inactivated status covers: G-M, G-D, E-G-M, E-G-D and E-LOH categories.

As with practically any treatment on target or off target toxicity is an unfortunate reality. Table 2.10 lists adverse events in small proof of concept clinical study (part of phase 1b study) of RG7112 in 20 patients with liposarcoma (Ray-Coquard et al., 2012).

Table 2.10 Adverse events in RG7112 liposarcoma study.

(Ray-Coquard, et al., 2012 Lancet Oncol. 2012 Nov;13(11):1133-40)

Adverse event	Total	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5
Nausea	14 (70%)	2 (10%)	11 (55%)	1 (5%)	0	0
Vomiting	11 (55%)	5 (25%)	4 (20%)	2 (10%)	0	0
Asthenia	9 (45%)	3 (15%)	6 (30%)	0	0	0
Diarrhoea	9 (45%)	7 (35%)	1 (5%)	1 (5%)	0	0
Thrombocytopenia	8 (40%)	1 (5%)	2 (10%)	2 (10%)	3 (15%)	0
Fatigue	6 (30%)	2 (10%)	4 (20%)	0	0	0
Neutropenia	6 (30%)	0	0	0	6 (30%)	0
Alopecia	4 (20%)	3 (15%)	1 (5%)	0	0	0
Constipation	3 (15%)	2 (10%)	1 (5%)	0	0	0
Decreased appetite	3 (15%)	2 (10%)	0	1 (5%)	0	0
Abdominal pain	2 (10%)	1 (5%)	1 (5%)	0	0	0
Anaemia	2 (10%)	1 (5%)	1 (5%)	0	0	0
Atrial fibrillation	2 (10%)	1 (5%)	1 (5%)	0	0	0
Back pain	2 (10%)	2 (10%)	0	0	0	0
Dysgeusia	2 (10%)	2 (10%)	0	0	0	0
Reflux	2 (10%)	1 (5%)	1 (5%)	0	0	0
Pain in extremity	2 (10%)	1 (5%)	1 (5%)	0	0	0
Pyrexia	2 (10%)	1 (5%)	1 (5%)	0	0	0
Urinary tract infection	2 (10%)	2 (10%)	0	0	0	0
Febrile neutropenia	1 (5%)	0	0	0	1 (5%)	0

Serious adverse events (grade four and above) such as thrombocytopenia, neutropenia and febrile neutropenia have been observed in that study in eight patients. Thrombocytopenia is a condition in which blood has a lower than normal number platelets. Platelets are essential for blood clot formation and therefore thrombocytopenia can lead to serious bleeding and can be potentially fatal. Neutropenia is a condition in which blood has a lower than normal number of neutrophils. Neutrophils are an essential part of the innate immune system and therefore neutropenia can lead to bacterial infections which could be life threatening. Febrile neutropenia is a serious condition in which neutropenia leads to infection and such condition can be potentially fatal. This highlights the importance of selecting only patients with wild type TP53 protein for treatment with MDM2 TP53 interaction inhibitor(s) in order to spare patients with inactivated TP53 from potential toxicity from treatments which have no chance to benefit them. And since as was previously discussed TP53 could be inactivated by different mechanisms it is important to account for multiple mechanisms of TP53 inactivation. Also from clinical and translational perspectives concentrating on patients with wild type TP53 protein may help to determine which patients may get the greatest therapeutic benefit from MDM2 TP53 interaction inhibitor(s).

2.4.2 BAP1 tumor suppressor status

Comprehensive analysis of tumor suppressor genes status can also be used to either discover or provide additional line of evidence for putative tumor suppressor genes. This could be illustrated using BRCA1-associated protein-1 (BAP1) as example. BAP1 was identified in 1998 as a novel ubiquitin hydrolase which binds to the BRCA1 RING finger and enhances BRCA1-mediated cell growth suppression and was proposed to be potential putative tumor suppressor gene (Jensen et al., 1998) (Ventii et al., 2008). Recent publications proposed the link between inactivation of BAP1, including germ line mutations, and different malignances, such as: lung mesothelioma, uveal melanoma, melanoma and renal cell carcinomas (Testa et al., 2011) (Carbone et al., 2012) (Peña-Llopis et al., 2012). The tumor suppressor status framework was run across CCLE for BAP1 gene and ten cell lines have been identified with BAP1 inactivation Table 2.11. In these ten cancer cell lines three different mechanisms of BAP1 inactivation have been identified: three cell lines have homozygous loss of function mutations, five cell lines have homozygous DNA loss and two cell lines have loss of one allele of BAP1 and loss of mRNA expression of other allele. Identification of multiple classical mechanisms of tumor suppressor inactivation in BAP1 gene increases the confidence of this gene belonging to tumor suppressor category.

Table 2.11 CCLE cancer cell lines with inactivated BAP1.

Cell Line Name	Cell Line Lineage	BAP1 status
TUHR14TKB	kidney carcinoma	G-M
VMRC-RCW	kidney carcinoma	G-M
HCC1187	breast ductal carcinoma	G-M
SK-BR-3	breast carcinoma	G-D
AU-565	breast carcinoma	G-D
IST-MES2	lung mesothelioma	G-D
NCI-H226	lung squamous cell carcinoma	G-D
NUGC-4	stomach signet ring adenocarcinoma	G-D
RT112/84	bladder carcinoma	E-G-D
JL-1	lung mesothelioma	E-G-D

TUHR14TKB, VMRC-RCW are renal cell carcinomas while IST-MES2, JL-1 are lung mesotheliomas. As mentioned above these two lineages have been potentially linked to inactivation of BAP1. Interestingly three CCLE breast carcinomas cell lines have inactivated BAP1. Since BAP1 is proposed to enhance BRCA1-mediated cell growth suppression it is possible to speculate that BAP1 can also be potentially inactivated in breast cancer and possibly in familial cases of breast cancer.

2.4.3 Potential future refinements

It is possible to envision how to further refine the introduced here methodology to comprehensively access status of tumor suppressor genes along the following lines.

The ability to accurately differentiate between two mutations affecting two different alleles and mutations affecting just one allele is limited by the length of sequencing reads. This limitation may have important implications. For example, in the case of APC, which is often disabled in colorectal cancers, a loss of function mutation affecting the first allele and also a different loss of function mutation affecting the second allele may occur relatively frequently. For example, in familial colorectal cancers one allele may be disabled by a germ line loss of function mutation while the other allele may be disabled by somatic loss of function mutation (Fodde, 2002). Therefore, in the case of APC, the total number of cell lines identified by the approach presented here as being genetically inactivated is likely to be underestimated. This situation is likely to improve with increase in length of sequencing reads, but in the meantime current estimates will represent minimum numbers of cell lines in such cases.

Genetic and epigenetic regulation of many tumor suppressors is complemented by posttranslational regulation. For example, TP53 has recently been referred to as one of the group of “massively regulated genes” with several alternative splicing sites and alternative translation initiation sites together generating potentially as many as ten distinct isoforms of the gene product (Hollstein and Hainaut, 2010). At the post-translational level, the biological activity of TP53 depends on its intracellular concentration and can be modulated by conformational changes, different intracellular localization, DNA-binding activity and interactions with other proteins. The accumulation and activity of the protein are also regulated by a suite of post-translational modifications that can include phosphorylation, acetylation, ubiquitination, sumoylation, neddylation, methylation and glycosylation. For example negative regulation of TP53 is provided in part by the MDM2 and MDM4 proteins, which are important determinants of TP53 abundance and subcellular localization.

Some of the tumor suppressors can be haploinsufficient (Payne and Kemp, 2005) and, in this case, the decrease in mRNA expression alone could have substantial tumor promoting effects.

It is interesting to note that recent work suggests that a loss of single copy of the chromosomes containing multiple tumor suppressor genes may lead to selective growth advantage (Xue et al., 2012) (Solimini et al., 2012). Future work on incorporating considerations of posttranslational regulations and haploinsufficiency may improve characterization of tumor suppressor genes status in cancer models.

Another interesting aspect of tumor suppressor genes biology is the existence of mutations which lead to partial, but not complete loss of function. There are several interesting examples of such cases. In TP53 the most frequent mutations lead to complete loss of function and in most cases also have a dominant negative effect, however 10% to 20% of low or medium frequency mutations belong to partial loss of function group (Petitjean et al., 2007). Classification of mutations as complete loss of function, partial loss of function or functional was mainly based on work in eight different yeast cells by examination of TP53 transactivation activity for all possible missense mutations produced by point mutagenesis across entire sequence of TP53 (Kato et al., 2003). As was previously mention in this chapter the TP53 transcriptionally activates or represses variety of target genes which in turn lead to number of different phenotypes including DNA repair, cell cycle arrest and apoptosis. Partial loss of function mutations in TP53 lead to mosaic phenotype with preservation or partial preservation of transactivation activity for some target genes and complete loss of transactivation activity for other target genes. Germline partial loss of function mutations of TP53 are also found in about 20% of Li-Fraumeni syndrome (LFS) and Li-Fraumeni-like syndrome (LFL) cases, and it seems like there is a potential relation between severely of TP53 mutations and on set of cancer (Olivier et al., 2003). Very intriguing clinical case of Familial Adenomatous Polyposis (FAP) is described by (Zajac et al., 2000), in this case there is a germline truncating mutation in APC and germline partial loss of function mutation in TP53. Combination of these two mutations seems to result in much more severe phenotype than typically observed with germline mutation in APC by itself.

The partial loss of function mutations are not unique to TP53 and are likely to be present in many tumor suppressor genes as has been reported for example in PTEN (Wu et al., 2000) , APC (Hughes et al., 2002) and RB1 (Sun et al., 2006). The existence and likely relevance of partial loss of function mutations in tumor suppressor genes suggests another potential future improvement for tumor suppressor status framework, it could be beneficial to create additional category for partial loss of function mutations.

The work on tumor suppressor status analysis also highlights the interesting possibility of identifying homozygous deletions based on exome sequencing data. More sensitive identification of homozygous deletions and especially small homozygous deletions would be useful addition to tumor suppressor status analysis and beyond. There are number of potential technical obstacles for such analysis, but it would be interesting to investigate the visibility of detection of small homozygous deletions using exome sequencing data.

There are many ways to take advantage of tumor suppressor status knowledge. For example, in studying the function of the gene in question it is often important to understand the aspects of the protein function due to its potential scaffolding function versus its enzymatic activity. Therefore cell line(s) with gene in question in G-D or E-G-D status would provide experimental model candidate(s) which likely have no protein and on another hand cell line(s) with gene in G-M status would provide experimental model candidate(s) with variety of different phenotypes including one in which protein is still present, but it's enzymatic activity is lost due to point mutation. Another interesting example could be combining tumor suppressor status with oncogene status in the pathway of interest should allow to better defining cell lines with dysregulated or un-affected pathway. Detailed examination of cell lines with multiple major tumor suppressors in the wild type state may help to better understand peculiarities of signaling cascades. From the translational perspective, preclinical stratification of anticancer drug sensitivity could benefit from systematically derived tumor suppressor status.

2.5 Summary

This chapter describes the approach used to incorporate several orthogonal genomic data types, such as mutation data, copy number, LOH and expression in order to account for different mechanisms of tumor suppressor genes inactivation mechanisms. Relationship between TP53 status and Nutlin-3 sensitivity was used to demonstrate the advantages of the integrative method, which allows accounting for multiple mechanisms of tumor suppressor genes inactivation, provides a more accurate assessment of tumor suppressor genes status than can be inferred by expression, copy number, or mutation alone. Relationship between RB1 status and PD-0332991 sensitivity was used to reiterate advantages of accounting for multiple mechanisms of tumor suppressor genes inactivation. Clinical relevance of the more accurate assessment of tumor suppressor genes status was highlighted with TP53 role in cancer and selection of potential treatment. BAP1 example was used to show how integrative tumor suppressor gene status can help to find new putative tumor suppressor genes. Potential future refinements were discussed as well.

Chapter Three: Gene sets activity analysis in Cancer Cell Line Encyclopedia

This chapter describes the details of the computational framework for the gene set activity analysis on sample by sample basis developed as part of thesis research and it is application to Cancer Cell Line Encyclopedia (CCLE).

Chapter 3.1 describes rationale for selecting gene sets for the analysis. Chapter 3.2.1 describes the approach used to generate gene set activity analysis on sample by sample basis. Chapters 3.2.2 and 3.2.3 take advantage of tissue specific genes and tissue specific processes respectively to validate analysis results. Chapter 3.2.4 discusses effects of permutation fractions depth on the analysis. Chapter 3.3.1 highlights results of gene set activity analysis on sample by sample basis across cancer cell lines of Cancer Cell Line Encyclopedia. Chapter 3.3.2 demonstrates utility of gene set activity analysis on sample by sample basis in generation of pSTAT5 mRNA expression signature. Chapter 3.3.3 shows interesting relationship between BRAF inhibitor sensitivity and MITF signaling discovered by gene set activity analysis on sample by sample basis. Chapter 3.4.1 highlights clinical relevance of pSTAT5 gene signature. Chapters 3.4.2 and 3.4.3 shows relationship between permutation fractions, gene set activity scores and Z-scores based statistics.

3.1 Selecting gene sets for analysis

One of the critical inputs into Gene Set Activity Analysis is collection of gene sets used for the analysis. Currently there are number of pathway databases providing access to collections of gene sets. Some of these pathway databases are freely publicly available and some are commercial and require the license. KEGG (Kanehisa et al., 2012) and BIOCARTA (http://www.biocarta.com) are the classical examples of freely publicly available pathway databases. GeneGo (www.genego.com), Ingenuity (www.ingenuity.com) and BIOBASE (www.biobase-international.com) are some of the commercially available pathway databases. Publications by (Bauer-Mehren et al., 2009) and (Ooi et al., 2010) provide in depth reviews of pathway databases.

MetaCore Data Base from GeneGo (www.genego.com) provides extensive collection of canonical and transcriptionally directional gene sets which went through additional rounds of manual curation. Curators at GeneGo review original literature sources and other available sources of information to assess the credibility of reported findings. These efforts are thought to improve the underlying quality of gene sets memberships and decrease the redundancy due to possible consolidation of gene sets from different sources. Another important advantage of GeneGo gene sets collection is an availability of directionality information for targets of transcription factors. This allows formation of transcriptionally directional gene sets which at least conceptually should allow the better read outs of transcriptional activity derived from mRNA profiling. Due to the above reasons the GeneGo gene sets were selected as primary source of gene sets for Gene Set Activity Analysis. The version of GeneGo gene sets used for the analysis consists of 566 canonical pathways and 716 transcriptionally directional gene sets, in total 1,282 GeneGo gene sets are selected for Gene Set Activity Analysis. Gene set names of GeneGo transcriptionally directional gene sets include an "inhibited" suffix which indicates that the gene set represents set of genes which are supposed to be transcriptionally inhibited by the transcription factor in question. Alternatively, genes with an "activated" suffix indicate that the gene set represents set of genes which are supposed to be transcriptionally activated by transcription factor in question.

The GeneGo gene sets collection is a very useful resource. However, since this is a commercial product, not every one may have an access to gene sets memberships. In order to address this point a subset of gene sets from the freely available Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) was also selected for Gene Set Activity Analysis. MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp) is a collection of annotated gene sets maintained by Broad Institute GSEA team. MSigDB gene sets collection is based mainly on online pathway databases and gene sets mined from publications in PubMed (www.ncbi.nlm.nih.gov/pubmed). Table 3.1 lists online pathway databases which are used as major sources of MSigDB Canonical Pathways. The names of MSigDB Canonical Pathways contain the following suffixes to indicate the source of the gene set: BIOCARTA, KEGG, REACTOME, SA (Sigma Aldrich), SIG (Signaling Gateway), ST (Signaling Transduction KE). The subset of MSigDB version 3.0 selected for the Gene Set Activity Analysis consist of 326 cytogenetic bands gene sets, 2120 gene sets representing signatures of genetic and chemical perturbations, 880 canonical pathways, 221 microRNA targets gene sets and 615 transcription factor targets gene sets. In total 4,162 MSigDB version 3.0 gene sets have been selected for the analysis.

Table 3.1 Online database sources for MSigDB Canonical Pathways.

Resource Name	URL link
BioCarta	http://www.biocarta.com
KEGG	http://www.genome.jp/kegg
Pathway Interaction Database	http://pid.nci.nih.gov
Reactome	http://www.reactome.org
SigmaAldrich	http://www.sigmaaldrich.com/life-science.html
Signaling Gateway	http://www.signaling-gateway.org
Signal Transduction KE	http://stke.sciencemag.org

3.2 Methods

3.2.1 Overview

Based on the literature review described in Chapter 1.3, a z-score based transformation approach was chosen for normalizing mRNA gene expression data. Gene set activity score calculations could be outlined as a following two-step process.

The first step in the process of calculating gene set activity scores is to perform z-score transformation for each probe set expression values across set of samples.

$$z_{i,j} = \frac{x_{i,j - \mu_i}}{\delta_i + \varepsilon}$$

$z_{i,j}$ is z-score value for probe set i in sample j

$x_{i,j}$ is MAS5 expression value for probe set i in sample j

μ_i is a mean of expression values for probe set i across all samples

δ_i is Standard Deviation of expression values for probe set i across all samples

ε is Standard Deviation Constant, 10 is used for MAS5 expression values

Standard Deviation Constant is used in order to prevent spikes in z-scores due to occasional artificially low Standard Deviation values.

The second step is to calculate gene set activity scores by adding up Zi,j score from genes in particular gene set and normalizing by square root of number genes in the gene set.

$$S_j \ = \frac{\sum_{i=1}^{N} Z_{i,j}}{\sqrt{N}}$$

S_j is the gene set activity score of the given gene set in sample *j*.

N - number of genes in the gene set.

Gene set activity scores could be calculated taking in the account expected directionality in gene expression of pathway genes.

$$S_j \ = \frac{\sum_{i=1}^{N} Z_{i,j} \, D_i}{\sqrt{N}}$$

where $D_i \ = \begin{cases} \ \ 1 \ \ if \ gene \ i \ is \ expected \ to \ be \ up \ regulated \\ -1 \ if \ gene \ i \ is \ expected \ to \ be \ down \ regulated \end{cases}$

Gene set activity scores also could be calculated using absolute values of Zi,j scores. This could be beneficial in cases than for example particular pathway consist of genes with corresponding Zi,j scores which cancel each other, however the level of signaling may still have biological consequences.

$$S_j \ = \frac{\sum_{i=1}^{N} |Z_{i,j}\,|}{\sqrt{N}}$$

Gene set scores permutation fractions are calculated based on gene set activity scores calculated for 1000 randomly generated gene sets of particular size in particular sample. Permutation fraction of 0.001 corresponds to unusually high activity of the gene set and permutation fraction of 1.000 corresponds to unusually low activity of the gene set. Since calculation of permutation fractions is a computationally expensive method the underlying software code was implemented in parallelizable way, so it could be run on a computational cluster. Permutation fractions could be converted to p-values by following approach:

1) If permutation fraction values are ≤ 0.5 they are multiplied by factor of two.

2) If permutation fraction values are ≥ 0.501 they are subtracted from 1.001 and multiplied by the factor of two.

3) If permutation fraction values are between 0.5 and 0.501, the p-values are set to one.

Before applying gene sets activity analysis to mRNA expression data set from Cancer Cell Line Encyclopedia it is important to show that the method described above produces biologically meaningful results. Two approaches have been used for this purpose.

3.2.2 Validation using tissue specific genes

The first approach is based on tissue specific genes which have been identified by Ge et al. (2005) for diverse collection of tissue types. Ge et al. (2005) identified tissue specific genes for 36 tissue types/subtypes, however some of them represent different anatomical areas of the brain and for the purpose of this analysis tissue specific genes for 29 tissue types were actually used. For many tissue types Ge et al. (2005) pooled RNA from multiple individual donors. Tissue specific genes are expected to be expressed in only/mostly one tissue type. Twenty nine gene sets have been created based on the data from Ge et al. (2005) each representing set of genes which supposed to be expressed only/mostly in corresponding normal human tissue.

Table 3.2 shows Gene Set Activity Scores for these 29 gene sets for liver sample GSM44702 which was used by Ge et al. (2005) in order to derive liver tissue specific genes in the first place.

Table 3.2 Gene Set Activity Scores for liver sample from the Ge et al. (2005) study.

Gene Set Name	Gene Set Size	Gene Set Activity Score	Permutation Fraction
Liver	171	54.4	0.001
Kidney	26	0.92	0.118
Small_Intestine	36	0.82	0.132
Fetal_Liver	21	0.7	0.15
Adrenal_gland	27	0.55	0.201
Spleen	28	0.52	0.228
Prostate	9	0.14	0.299
Ovary	8	0.14	0.309
Colon	14	0.01	0.407
Thyroid	17	-0.03	0.42
Lung	32	-0.13	0.524
Trachea	12	-0.14	0.469
Placenta	54	-0.3	0.668
Salivary_gland	17	-0.36	0.657
Bladder	11	-0.47	0.738
Uterus	9	-0.5	0.756
Pancreas	39	-0.6	0.805
Heart	23	-0.74	0.861
Thymus	41	-0.9	0.927
Bone_Marrow	38	-0.98	0.942
Breast	14	-1.04	0.971
Fetal_Brain	33	-1.15	0.968
Skin	73	-1.15	0.955
Stomach	25	-1.16	0.972
Skeletal_Muscle	71	-1.26	0.979
Testis	376	-1.57	0.996
Fetal_Lung	37	-1.57	1
Brain	299	-6.56	1

As expected, the Gene Set Activity Score is clearly much higher for the gene set representing liver-specific genes than for other 28 gene sets representing other tissue types. Also as expected the permutation fraction for the gene set representing liver-specific genes is much lower than for the remaining 28 gene sets representing other tissue types.

Table 3.3 Top liver specific Gene Set Activity Scores across set of normal samples.

Sample Name	Gene Set Activity Score	Permutation Fraction
normal liver GSM144285	89.47	0.001
normal liver GSM18954	79.32	0.001
normal liver GSM44702	77.19	0.001
normal liver GSM144282	73.69	0.001
normal liver GSM18953	69.01	0.001
normal liver GSM18906	26.66	0.001
normal liver GSM18905	22.52	0.001
normal liver GSM44706	18.67	0.001
normal pancreas GSM26415	9.67	0.025
normal pancreas GSM26419	9.46	0.109
normal pancreas GSM26423	8.95	0.023
normal pancreas GSM26428	8.80	0.145
normal autonomic ganglia GSM19011	7.65	0.014
normal pancreas GSM26302	7.45	0.189
normal autonomic ganglia GSM19012	7.32	0.005
normal kidney GSM18955	7.28	0.001
normal striated muscle GSM19014	7.26	0.001
normal pancreas GSM26399	7.05	0.027
normal upper aero-digestive tract GSM227859	7.01	0.02
normal striated muscle GSM19008	6.91	0.003
normal nerve sheath GSM19005	6.65	0.013
normal pancreas GSM26416	6.62	0.227
normal kidney GSM146789	6.57	0.001
normal striated muscle GSM19013	6.56	0.033
normal pancreas GSM26296	6.56	0.004
normal pancreas GSM26424	6.43	0.108
normal upper aero-digestive tract GSM227858	6.37	0.03
normal nerve sheath GSM19006	6.35	0.014
normal pancreas GSM26407	6.34	0.059
normal kidney GSM12098	6.17	0.001

In order to take a look at Gene Set Activity Scores for tissue specific gene sets in the independent set of normal human tissues, gene sets activity analysis for 29 tissue specific gene sets was run for 938 publicly available normal human tissue samples. Table 3.3 shows Gene Set Activity Scores for the liver specific gene set in 30 samples with highest Gene Set Activity Scores. The top entries in this table are enriched for liver samples and liver samples in general have noticeably higher Gene Set Activity Scores.

The human liver is heavily involved in numerous metabolic related processes, some of which only take place in liver and not in any other organs; partially due to these reasons the liver-specific gene set is relatively large, consisting of 171 member genes. Because of this it would be prudent to also examine a different tissue specific gene set of smaller size. The ovary-specific gene is an appropriate example for this purpose as it is the smallest tissue specific gene set, consisting of only 8 member genes. Table 3.4 shows Gene Set Activity Scores for 29 gene sets for ovary sample GSM44674 which was used by (Ge et al., 2005) in order to derive liver- specific genes in the first place. As expected Gene Set Activity Score is clearly much higher for gene set representing ovary tissue specific genes than for other 28 gene sets representing other tissue types. Also as expected the permutation fraction for the gene set representing ovary-specific genes is lower than for most others in the 28 gene sets representing other tissue types. In order to take a look at Gene Set Activity Scores for ovary-specific gene sets in the independent set of normal human tissues, gene sets activity analysis of 29 tissue specific gene sets calculated for 938 publicly available normal human tissue samples was used one more time.

Table 3.4 Gene Set Activity Scores for ovary sample from (Ge et al., 2005) study.

Gene Set Name	Gene Set Size	Gene Set Activity Score	Permutation Fraction
Ovary	8	10.54	0.001
Fetal_Lung	37	2.05	0.001
Bladder	11	0.74	0.063
Prostate	9	0.27	0.216
Uterus	9	0.21	0.255
Heart	23	-0.15	0.566
Adrenal_gland	27	-0.27	0.684
Thyroid	17	-0.52	0.861
Salivary_gland	17	-0.55	0.871
Breast	14	-0.63	0.931
Skeletal_Muscle	71	-0.68	0.909
Kidney	26	-0.74	0.969
Stomach	25	-0.85	0.981
Fetal_Brain	33	-0.86	0.981
Lung	32	-0.88	0.984
Colon	14	-0.92	0.989
Placenta	54	-0.98	0.988
Trachea	12	-0.98	0.99
Pancreas	39	-1.02	0.994
Fetal_Liver	21	-1.16	0.999
Spleen	28	-1.22	1
Bone_Marrow	38	-1.24	1
Small_Intestine	36	-1.46	1
Skin	73	-1.48	1
Thymus	41	-1.6	1
Testis	376	-1.78	1
Liver	171	-2.51	1
Brain	299	-5.77	1

Table 3.5 shows Gene Set Activity Scores for ovary specific gene set in 30 samples with highest Gene Set Activity Scores. The top entries in this table are enriched for ovary samples and ovary samples in general have noticeably higher Gene Set Activity Scores.

Table 3.5 Top ovary specific Gene Set Activity Scores across set of normal samples.

Sample Name	Gene Set Activity Score	Permutation Fraction
normal ovary GSM44674	20.36	0.001
normal ovary GSM139478	11.39	0.001
normal ovary GSM139477	11.23	0.001
normal ovary GSM139476	9.9	0.001
normal ovary GSM139479	8.75	0.001
normal pancreas GSM26415	4.79	0.043
normal striated_muscle GSM19008	4.67	0.005
normal pancreas GSM26419	4.35	0.079
normal haematopoietic_and_lymphoid_tissue GSM30583	3.12	0.014
normal ovary GSM18998	2.93	0.013
normal pancreas GSM26399	2.81	0.066
normal pituitary GSM44699	2.58	0.005
normal upper_aerodigestive_tract GSM227859	2.56	0.075
normal autonomic_ganglia GSM19011	2.5	0.094
normal placenta GSM18968	2.39	0.026
normal striated_muscle GSM19007	2.27	0.093
normal stomach GSM144271	2.26	0.016
normal pancreas GSM26423	2.24	0.18
normal haematopoietic_and_lymphoid_tissue GSM30584	2.23	0.078
normal stomach GSM144266	2.19	0.017
normal ovary GSM18997	2.18	0.034
normal adrenal_gland GSM18995	2.18	0.028
normal stomach GSM144272	2.16	0.018
normal smooth_muscle GSM18959	2.11	0.003
normal pancreas GSM26416	2.11	0.194
normal striated_muscle GSM19014	2.01	0.113
normal kidney GSM52495	1.98	0.006
normal pancreas GSM26424	1.97	0.178
normal pituitary GSM19021	1.89	0.02
normal adrenal_gland GSM18948	1.86	0.029

3.2.3 Validation using tissue specific processes

The second approach is based on tissue specific processes, for example Bile Acid Biosynthesis which is a liver-specific process. Unfortunately this approach is somewhat limited due to the lack of authoritative and comprehensive sources for tissue specific processes. Nevertheless this could be an informative exercise especially for such tissues as liver which should have gene expression data from genes involved in number of metabolic related process. As mention in Chapter 3.1 MetaCore Data Base from GeneGo (www.genego.com) provides an extensive collection of manually reviewed canonical and transcriptionally directional gene sets, the version used for this analysis consist of about 1,282 gene sets. Table 3.6 shows Gene Set Activity Scores for these gene sets in liver sample GSM44702 from (Ge et al., 2005) with highest Gene Set Activity Scores. As expected the top entries in this table are enriched for metabolic related process including Bile Acid Biosynthesis and Estradiol metabolism which are considered to be liver-specific processes. Also the top entry in the table is HNF4A-activated gene set which contains genes transcriptionally activated by Hepatocyte nuclear factor 4 alpha (HNF4A) transcriptional factor. HNF4A transcriptionally activates Hepatocyte nuclear factor 1 alpha (HNF1A) which is liver specific transcriptional factor. On the other hand Appendix A1.4 shows Gene Set Activity Scores for gene sets in the same liver sample GSM44702 from (Ge et al., 2005) with the lowest Gene Set Activity Scores. As expected the gene sets with lowest Gene Set Activity Scores seem to be enriched for entries which are not metabolic or liver-specific processes.

Table 3.6 Top GeneGo Gene Set Activity Scores for liver sample from (Ge et al., 2005) study.

Gene Set Name	Gene Set Size	Gene Set Activity Score	Permutation Fraction
HNF4A_activated	79	18.90	0.001
Androstenedione and testosterone biosynthesis and metabolism p.2	17	14.05	0.001
Immune response _Lectin Induced complement pathway	39	13.67	0.001
Estradiol metabolism	21	13.17	0.001
HNF1A_activated	57	13.14	0.001
NR1I3_activated	16	12.57	0.001
HNF1B_HNF1A_activated	34	12.38	0.001
Retinol metabolism	31	12.27	0.001
Immune response _Classic complement pathway	42	12.24	0.001
PXR_activated	37	11.74	0.001
DBP_activated	11	11.67	0.001
Bile Acid Biosynthesis	22	11.58	0.001
CEBPA_activated	95	11.38	0.001
Glycine, serine, cysteine and threonine metabolism	45	11.12	0.001
Acetaminophen metabolism	15	10.96	0.001
Immune response _Alternative complement pathway	23	10.91	0.001
STAT3_activated	87	10.85	0.001
2-Naphthylamine and 2-Nitronaphtalene metabolism	19	10.78	0.001
Leucune, isoleucine and valine metabolism.p.2	25	10.41	0.001
Androstenedione and testosterone biosynthesis and metabolism p.3	16	10.30	0.001
CEBPB_activated	114	10.19	0.001
1-Naphthylamine and 1-Nitronaphtalene metabolism	12	9.99	0.001
Alanine, cysteine, and L-methionine metabolism	23	9.37	0.001
GCR_activated	59	9.28	0.001
Tryptophan metabolism	32	9.27	0.001
Benzo[a]pyrene metabolism	12	9.24	0.001
SHP_inhibited	11	9.06	0.001
HNF3B_activated	38	8.80	0.001
Vitamin E (alfa-tocopherol) metabolism	17	8.65	0.001
Androstenedione and testosterone biosynthesis and metabolism p.1	19	8.42	0.001

Taken together the above examples show that the implementation of the method for Gene Set Activity Analysis on individual samples level introduced at the beginning of the methods section produces the result which seems to agree with biological intuition.

3.2.4 Permutation fractions depth

The permutation fractions in the above method are calculated with 1,000 permutations which in most cases is sufficiently deep and at the same time computationally manageable. However since the number of permutations is finite in some cases permutation fractions by themselves do not provide an ideal level of resolution. For example: Table 3.7 shows Gene Set Activity Scores for 29 tissue specific gene sets for spleen sample GSM44673 which was used by (Ge et al., 2005) in order to derive spleen specific genes in the first place. As expected the Gene Set Activity Score is clearly much higher for the gene set representing spleen tissue specific genes than for the other 28 gene sets representing other tissue types. However the top three tissue specific gene sets spleen, thymus and bone-marrow have the same permutation fraction of 0.001 which is likely due to an insufficient number of permutations to differentiate between them. Table 3.7 also shows permutation fractions calculated with 10,000 and 100,000 permutations. As it can be seen from the table, an increase to 10,000 permutations allows differentiating permutation fractions between thymus and bone-marrow specific gene sets; and an increase to 100,000 permutations allows differentiating permutation fractions between spleen, thymus and bone-marrow specific gene sets.

Table 3.7 Gene Set Activity Scores for spleen sample from (Ge et al., 2005) study.

Gene Set Name	Gene Set Size	Gene Set Activity Score	Permutation Fraction 1000 permutations	Permutation Fraction 10,000 permutations	Permutation Fraction 100,000 permutations
Spleen	28	20.31	0.001	1.00E-04	1.00E-05
Thymus	41	2.76	0.001	1.00E-04	5.00E-05
Bone Marrow	38	2.1	0.001	0.0011	8.30E-04
Thyroid	17	0.78	0.088	0.0536	0.05388
Bladder	11	0.1	0.401	0.5113	0.51137
Breast	14	-0.03	0.511	0.3957	0.39837
Ovary	8	-0.13	0.599	0.6805	0.68566
Uterus	9	-0.16	0.651	0.5208	0.51805
Lung	32	-0.17	0.663	0.7227	0.73272
Placenta	54	-0.31	0.767	0.9094	0.91576
Adrenal gland	27	-0.47	0.886	0.9265	0.928
Trachea	12	-0.5	0.874	0.8504	0.84559
Prostate	9	-0.52	0.924	0.8924	0.88959
Fetal Lung	37	-0.53	0.91	0.9035	0.90895
Salivary gland	17	-0.65	0.962	0.932	0.93004
Kidney	26	-0.69	0.964	0.9655	0.96621
Stomach	25	-0.69	0.958	0.8638	0.86359
Pancreas	39	-0.76	0.98	0.9787	0.97568
Colon	14	-0.79	0.982	0.9853	0.98521
Heart	23	-0.79	0.982	0.9728	0.97248
Fetal Liver	21	-0.83	0.98	0.9719	0.97203
Skeletal Muscle	71	-0.9	0.991	0.9696	0.9698
Small Intestine	36	-0.91	0.993	0.9898	0.99092
Fetal Brain	33	-1.02	0.995	0.9864	0.98679
Skin	73	-1.3	1	0.9999	0.99983
Testis	376	-1.35	1	0.9998	0.99975
Liver	171	-2.04	1	1	1
Brain	299	-5.33	1	1	1

3.3 Results

3.3.1 Overview

In order to get a comprehensive view of gene sets activity across CCLE each of the canonical and transcriptionally directional gene sets available from GeneGo Inc. and in addition each tissue specific gene set described in the above Methods section was used to generate activity scores and permutation fraction for 964 CCLE cancer cell lines for which mRNA expression profiles were available. As was mention in section 3.1 there are 1,282 gene sets from GeneGo and 29 tissue specific gene sets adding up to a total number of 1,311 gene sets. Appendix A2 CD contains files with gene sets activity scores and permutation fractions for GeneGo canonical gene sets, GeneGo transcriptionally directional gene sets, tissue specific gene sets across 964 CCLE cancer cell lines. The results of this analysis were used along with other CCLE data sets as inputs in (Barretina et al., 2012) for modeling anticancer drug sensitivity. One of the reasons for using Gene Set Activity Analysis output for modeling anticancer drug sensitivity has to do with its potential to help with the decreasing multidimensionality of data input into machine learning algorithms.

Cancer cell line mRNA expression profiles are well known to cluster in such way that solid and hematopoietic cancers lines tend to cluster away from each other while cell lines from the same cancer indications in general cluster together (Armstrong et al., 2002). To access Gene Set Activity Scores generated for the CCLE the unsupervised hierarchical clustering with correlation as distance measure was used to cluster Gene Set Activity Scores for 964 CCLE lines.

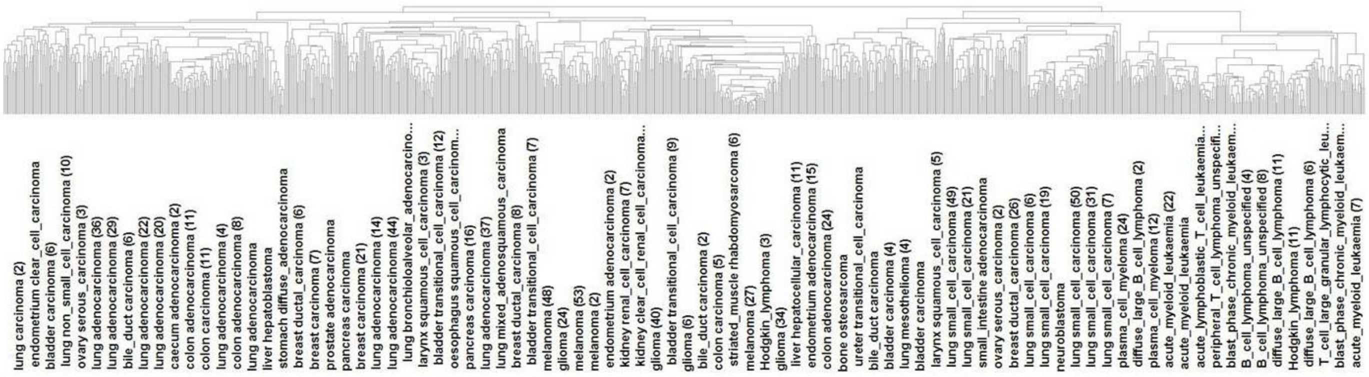


Figure 3.1 Hierarchical clustering of Gene Set Activity Scores for CCLE lines.

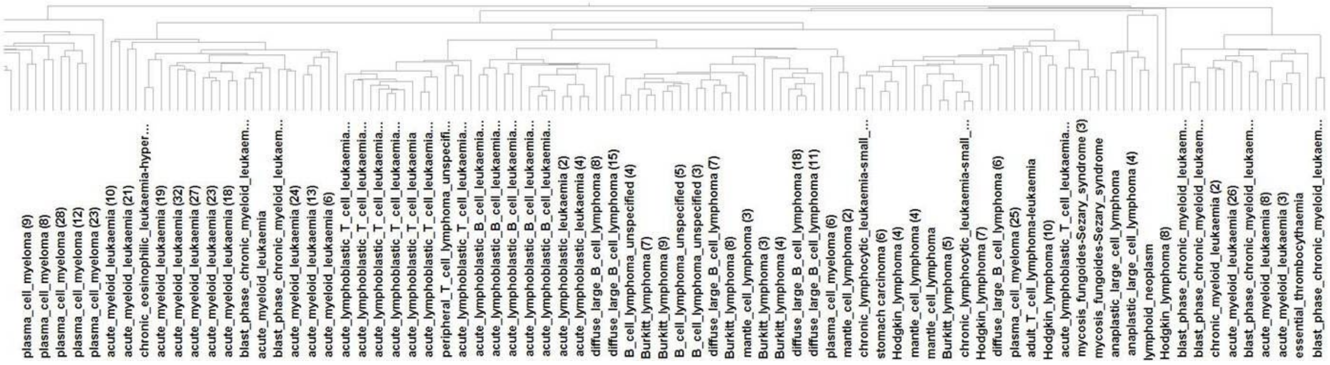


Figure 3.2 Hierarchical clustering of Gene Set Activity Scores for hematopoietic CCLE lines.

Figure 3.1 shows hematopoietic cancer cell lines clustering in the right corner and solid cancer cell lines to the left of the hematopoietic cluster. Figure 3.2 shows in more details the hematopoietic cancer cell lines cluster, this more detailed view shows that cell lines from the same indications in general cluster together, for example we can observed the tendency of plasma cell myelomas, acute myeloid leukaemias and acute lymphoblastic leukaemias to form separate clusters. Therefore Figures 3.1 and 3.2 seems to suggest that Gene Set Activity Scores preserved underlying mRNA expression profiles data structure while at the same time helping to decrease dimensionality from about 22,000 features to about 1,300 features.

Also as was mentioned in section 3.1, a subset of MSigDB version 3.0 gene sets collection was also selected as source of gene sets for Gene Set Activity Analysis. In total gene sets activity scores and permutation fractions were calculated for 4,162 MSigDB version 3.0 gene sets. Appendix A2 CD contains files with gene sets activity scores and permutation fractions for MSigDB cytogenetic bands gene sets, MSigDB gene sets representing signatures of genetic and chemical perturbations, MSigDB canonical pathways, MSigDB microRNA targets gene sets and MSigDB transcription factor targets gene sets across 964 CCLE cancer cell lines.

3.3.2 pSTAT5 mRNA expression signature

3.3.2.1 Gene signature enrichment

As was mentioned in section 1.4 deriving a gene expression based signature indicative of STAT5 activation status could be of practical importance. STAT5 is a transcriptional factor for which there is reasonably well defined set of transcriptional targets (MetaCore from GeneGo Inc.). The 47 genes which are considered to be transcriptional targets of STAT5 and have probe sets on the U133Plus2 array were selected for analysis. For each of the 47 genes, the best probe set was chosen based on combination of manual review, a computational approach (Nurtdinov et al., 2010) and also by taking in the account dynamic expression range across all CCLE cancer cell lines.

Two sets of CCLE hematopoietic cell lines with pSTAT5 Western blot data were available for analysis. The first set has data for 28 cell lines with 8 pSTAT5 positive and 20 pSTAT5 negative by Western blotting (Table 3.8). The second set has data for 12 unique cell lines, with 6 pSTAT5 positive and 6 pSTAT5 negative by Western blotting (Table 3.8). Taken together, this means that 40 unique cell lines with known pSTAT5 status are available for analysis (Table 3.8).

Table 3.8 pSTAT5 status for 40 CCLE hematopoietic cell lines

Cell Line Name	pSTAT5 set 1	pSTAT5 set 2
THP-1	N	N
PL-21		N
OCI-AML2		N
NOMO-1		N
HL-60		N
KASUMI-1		N
SKM-1		N
MM1-S	N	
ST486	N	
NCI-H929	N	
JM1	N	
Loucy	N	
RPMI 8226	N	
Toledo	N	
MC116	N	
Reh	N	
KMS-12-BM	N	
RS4;11	N	
BDCM	N	
U-937	N	
HD-MY-Z	N	
HuNS1	N	
SUP-T1	N	
CA46	N	
RL	N	
HH	N	
MOLM-13	Y	Y
AML-193	Y	Y
Set-2	Y	Y
TF-1	Y	Y
HEL 92.1.7	Y	Y
EOL-1		Y
F-36P		Y
Kasumi-6		Y
MV-4-11		Y
M-07e		Y
OCI-AML5		Y
K-562	Y	
SUP-B15	Y	
MEG-01	Y	

Gene set activity scores were calculated for 40 cell lines from Table 3.8 using a gene set consisting of 47 STAT5 transcriptional targets. Figure 3.3 shows that there is a tendency for pSTAT5 negative cell lines to have lower gene set activity scores and for pSTAT5 positive cell lines have a higher gene set activity scores. Based on this encouraging pattern data from the first set of cell lines were used for signature enrichment and data from second set of cell lines were used for signature validation.

In order to try to improve gene signature, fold changes and Student's t-Test probabilities between pSTAT5 positive and pSTAT5 negative cell lines were calculated using data from the enrichment cell line set. For fold change calculations, a value of 50 was added to the expression averages for pSTAT5 positive and pSTAT5 negative cell lines in order to decrease noise from low expressing genes. Positive values indicate higher expression in pSTAT5 positive lines, while negative values indicate higher expression in pSTAT5 negative lines. Student's t-Test was run using two-tailed distribution and homoscedastic settings. Appendix A1.5 provides the results for all 47 genes.

Data from Appendix A1.5 was used to create 3 genes sets (Table 3.9). The first one included four genes (PIM1, CISH, SOCS2, ID1) with lowest p-values and fold changes above 4. The second gene set contains the aforementioned four genes and LCN2 and EPOR, both of which have fold changes around 2 and p-values below 0.01. The third gene set carries the additional gene, EGR1, which has a fold change of around 2.7, but a p-value of ~ 0.06. Table 3.10 provides the gene set activity scores for three gene signatures across all cell lines.

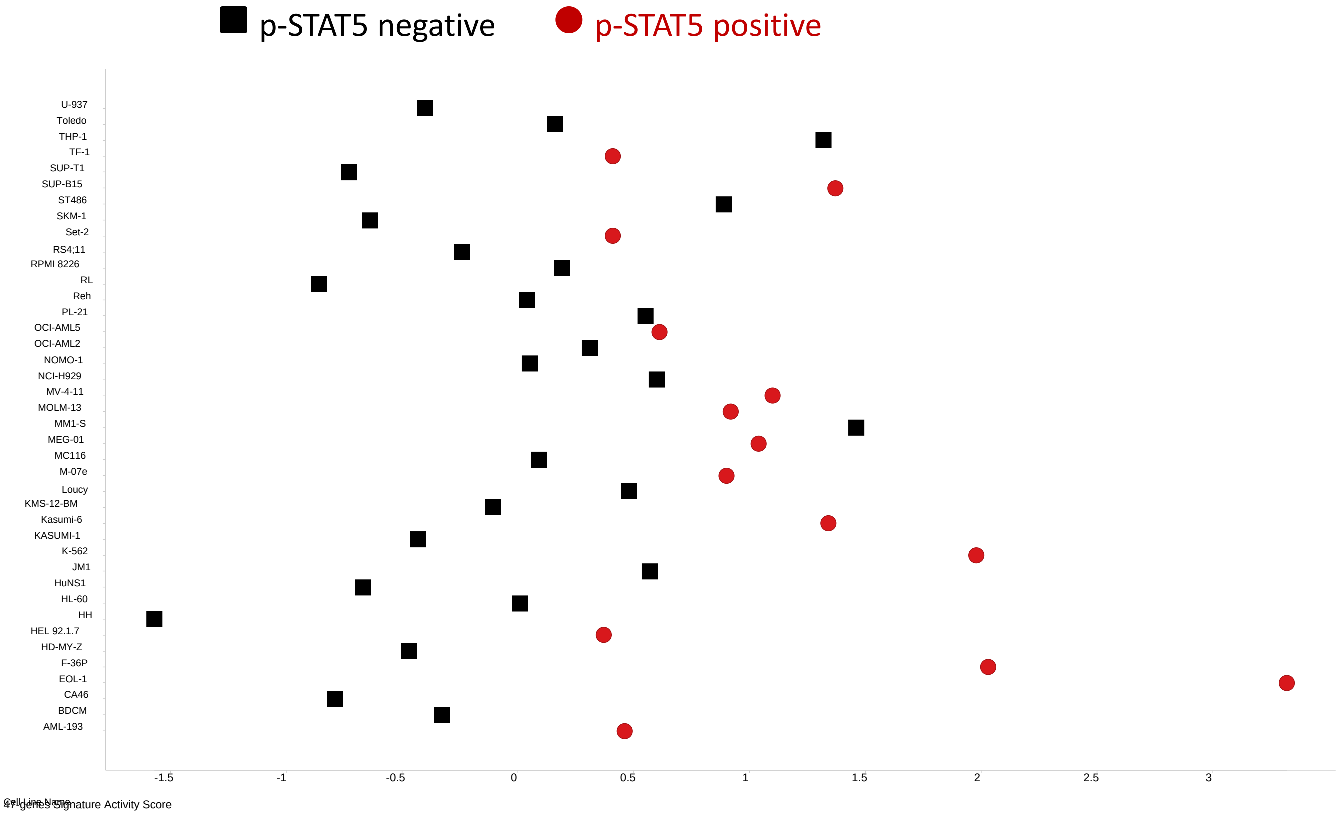


Figure 3.3 pSTAT5 status and gene signature scores for a 47-gene signature

Table 3.9 Putative gene signatures to differentiate pSTAT5 status

4-gene signature	6-gene signature	7-gene signature
PIM 1	PIM 1	PIM 1
CISH	CISH	CISH
SOCS2	SOCS2	SOCS2
ID1	ID1	ID1
	LCN2	LCN2
	EPOR	EPOR
		EGR1

Table 3.10 Gene signature activity scores for three gene signatures across 40 cell lines

Cell Line Name	4-genes	6-genes	7-genes	pSTAT5 set 1	pSTAT5 set2	pSTAT5
THP-1	-0.82	-0.88	-0.93	N		N
PL-21	-0.72	-0.41	-0.22		N	N
OCI-AML2	0.43	0.3	-0.07		N	N
NOMO-1	0.12	-0.06	-0.37		N	N
HL-60	-0.83	-0.92	-1.17		N	N
KASUMI-1	-0.56	-0.7	-0.95		N	N
SKM-1	-0.88	-0.94	-1.2		N	N
MM1-S	-0.68	-0.61	-0.9	N		N
ST486	-1.02	-0.99	-1.25	N		N
NCI-H929	-0.6	-0.55	-0.68	N		N
JM1	-0.99	-1.1	-1.38	N		N
Loucy	-1.03	-1	-1.27	N		N
RPMI 8226	-0.71	-0.73	-1.02	N		N
Toledo	-0.98	-1.11	-1.39	N		N
MC116	-1.06	-1.01	-1.26	N		N
Reh	0.14	0	-0.38	N		N
KMS-12-BM	-0.16	-0.04	-0.41	N		N
RS4;11	-0.7	-0.85	-1.12	N		N
BDCM	-0.87	-0.94	-1.05	N		N
U-937	-0.48	-0.5	-0.81	N		N
HD-MY-Z	-0.85	-0.74	-0.32	N		N
HuNS1	-0.76	-0.71	-1.01	N		N
SUP-T1	-0.89	-0.92	-1.19	N		N
CA46	-0.94	-0.98	-1.25	N		N
RL	-1.12	-1.13	-1.41	N		N
HH	-1.01	-0.98	-1.27	N		N
MOLM-13	2.13	1.79	1.36	Y		Y
AML-193	2.46	1.74	1.32	Y		Y
Set-2	1.72	2.38	1.93	Y		Y
TF-1	1.65	2.63	2.07	Y		Y
HEL 92.1.7	1.7	1.38	1.42	Y		Y
EOL-1	7.46	5.98	5.22		Y	Y
F-36P	4.32	4.55	3.93		Y	Y
Kasumi-6	2.47	1.77	1.36		Y	Y
MV-4-11	0.81	0.66	0.37		Y	Y
M-07e	3.06	2.34	1.99		Y	Y
OCI-AML5	1	0.64	0.24		Y	Y
K-562	6.12	4.92	4.63	Y		Y
SUP-B15	1.21	0.69	0.4	Y		Y
MEG-01	3.09	2.94	2.53	Y		Y

3.3.2.2 Gene signature validation

The validation set of cell lines was used to independently validate these gene sets. For the three gene signatures, the probability associated with the Student's t-Test between gene signature activity scores for pSTAT5 positive and pSTAT5 negative cell lines was calculated using data from independent validation cell lines set and in all cell lines from enrichment and validation sets combined. Student's t-Test was run using two-tailed distribution and heteroscedastic settings. Table 3.11 provides the results for three gene signatures in the validation set cell lines and in all cell lines. As can be seen from Table 3.11, all three gene signatures have p-values below 0.05 in the independent validation set. The lowest p-value is observed for the 7-gene signature in cell lines of set 1 and set 2 combined.

Table 3.11 t-test p-values for three gene signatures across 40 cell lines

Cell lines set	4-genes signature t-test p-value	6-genes signature t-test p-value	7-genes signature t-test p-value
Validation (set 2)	0.015	0.016	0.016
Enrichment + validation (set 1 and set 2)	1.305E-05	6.001E-06	5.511E-06

Figure 3.4 shows the relationship between the pSTAT5 status and 4-gene signature activity scores across all cell lines; this demonstrates the ability of the signature to discriminate between pSTAT5 positive and pSTAT5 negative haematopoietic cell lines.

Interestingly, in 4 genes signature two of the genes (*PIM1* and *ID1*) are known positive activators of the JAK-STAT5 pathway, and the other two (*SOCS2* and *CISH*) are known negative regulators thereof.

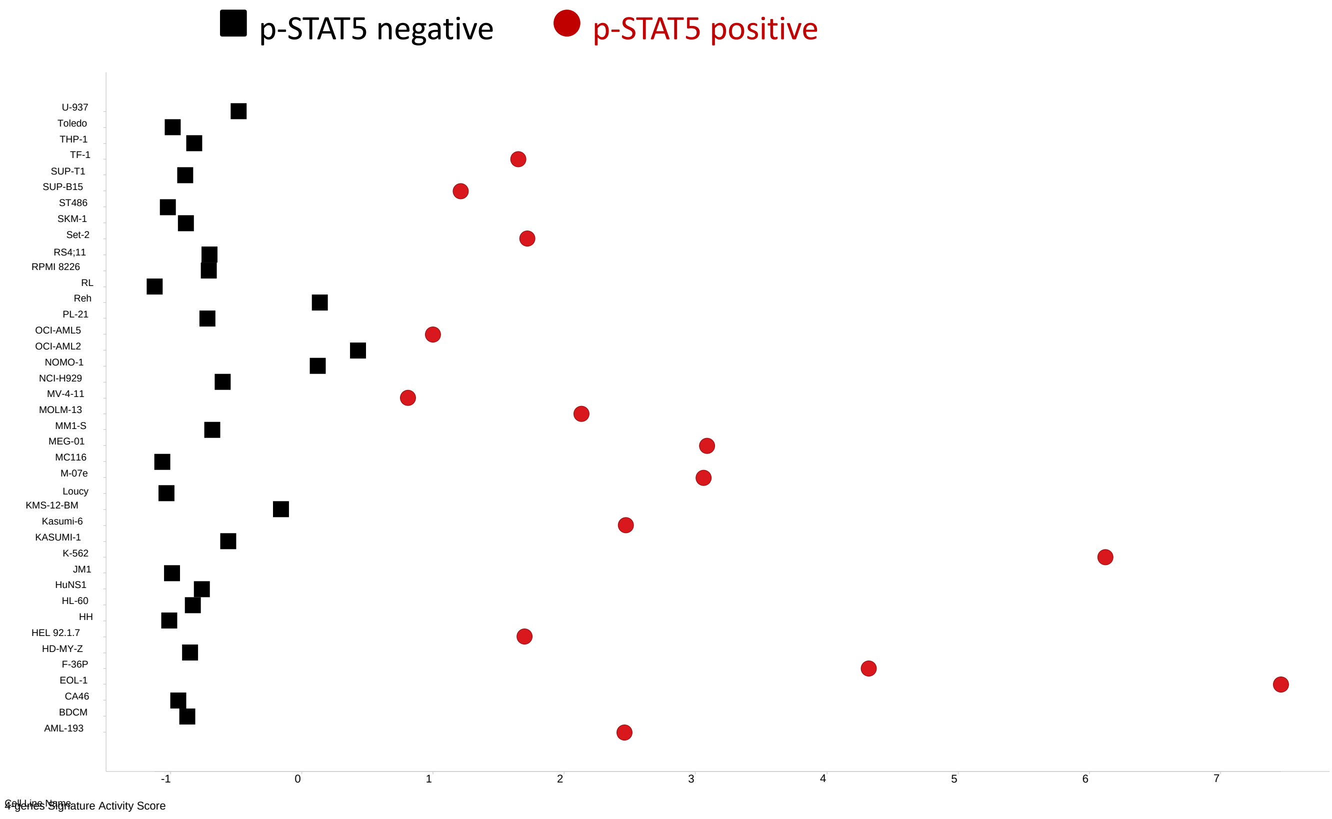


Figure 3.4 pSTAT5 status and gene signature scores for a 4-gene signature

Pim kinases, which play critical roles in cell survival and are believed to be important targets in hematological malignancy, are reportedly regulated at the transcriptional level, immediately downstream of STAT signaling (Amaravadi and Thompson, 2005) (Bachmann and Möröy, 2005) (Wang et al., 2001) (White, 2003). Indeed, it has been reported that JAK inhibition would lead to the down modulation of PIM1 and PIM2 mRNA in SET-2 cells (Gozgit et al., 2008) (Wood et al., 2009). Thus, activated JAK2 would result in increased PIM1 mRNA via the activation of STAT5.

ID1 is involved in cell cycle progression and hematopoietic stem cell self-renewal (Norton et al., 1998) (Jankovic et al., 2007) (Perry et al., 2007). Studies have shown that ID1 promotes survival during erythroid differentiation in fetal liver (Wood et al., 2009). ID1 is a target of the JAK2-STAT5 signaling pathway in erythroid cells with its expression being modulated by activated STAT5. Furthermore, ID1 transcript levels are elevated in cells harboring the JAK2V617F mutation, when compared to JAK2 WT cells, and they are down modulated by a JAK inhibitor (Perry et al., 2007).

SOCS2 is a member of the suppressors of cytokine signal (SOCS) family (Lai SY et al., 2010). SOCS members inhibit JAK kinase activity and facilitate proteosomal degradation of JAK (Wood et al., 2009). It has been reported that SOCS expression is correlated with STAT5 activity (Sen et al., 2012).

Cytokine inducible SH2-domain containing protein (CISH) is an immediate early gene induced by IL-2, IL-3, EPO and GM-CSF (Endo et al., 1997). It is a STAT5 target gene, a member of the SOCS family and it inhibits JAK-STAT5 signaling via a feedback loop (Matsumoto et al., 1997) (Ram and Waxman, 1999). Specifically, CISH expression is increased with the activation of JAK-STAT5 in the presence of specific ligands such as IL-2 or Prolactin (PRL) in respective model systems (Fang et al., 2008). Additionally, IL-2 induced CISH expression was shown to be suppressed by a dominant-negative STAT5 (Mitchell et al., 2003).

3.3.2.3 Gene signature and pharmacodynamics response to ruxolitinib

The functional relevance of these 4 genes to pSTAT5 increases the validity of using such a signature to predict pathway activation. In order to further validate 4 genes signature my collaborators on manuscript (which is under review) performed set of wet lab experiments outlined in this section (Appendix A1.1, Sonkin et al.).

Nine different cell-lines (Table 3.12) were treated with two concentrations of the JAK1/2 inhibitor ruxolitinib (0.2µM and 1µM) and DMSO, and sampled at two functional time points (4 and 24 hrs).

Table 3.12 List of cell lines used for additional validation of pSTAT5 signature

Cell Line Name	pSTAT5	JAK2 mutation
AML-193	Y	NO
HEL 92.1.7	Y	YES
Set-2	Y	YES
TF-1	Y	NO
UKE-1	Y	YES
PL-21	N	NO
Reh	N	NO
RPMI 8226	N	NO
U-937	N	NO

For each treatment and time point the total STAT5 and pSTAT5 levels were examined by Western blot analysis and the expression of the four signature genes was determined by qPCR. The baseline expression levels of the pSTAT5 positive cell lines were notably higher than that of the pSTAT5 negative cell lines (Figure 3.5). In each case there was at least a one ΔCT increase between the highest negative and lowest positive cell line level of expression, with most having large expression differences. If the signature genes are examined cumulatively the difference is quite large.

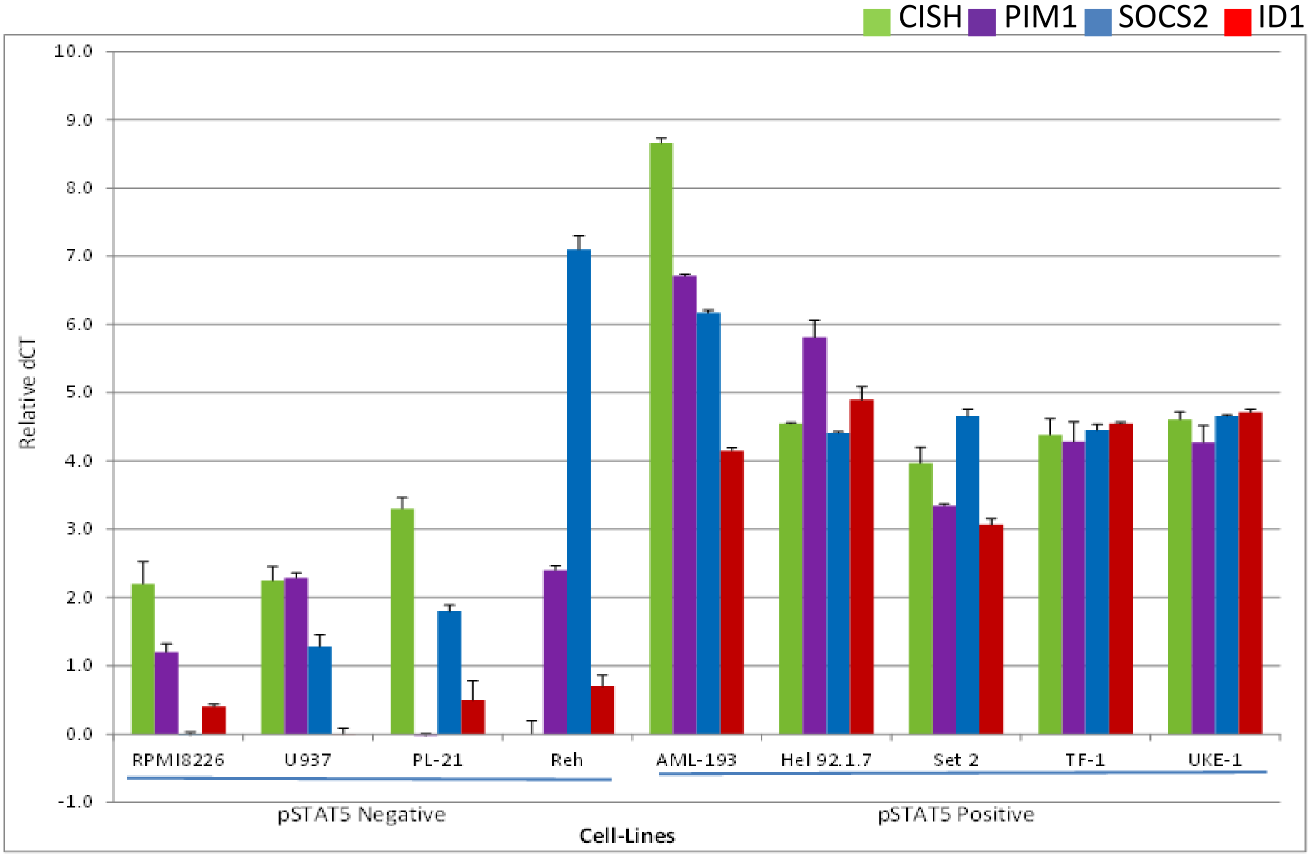


Figure 3.5 qRT-PCR results after 4 hours of treatment with vehicle (DMSO)

Figure 3.6 demonstrates modulation of pSTAT5 and the expression of the 4 signature genes by ruxolitinib in two pSTAT5 positive cell lines and two pSTAT5 negative cell lines. Figure 3.7 shows normalized average expression in 5 pSTAT5 positive cell lines and 4 pSTAT5 negative cell lines in response to ruxolitinib, normalized to the respective DMSO-treated controls. The average difference in the 4 and 24 hour treatments in the pSTAT5 positive cell lines showed a reduction in the 4 signature genes at both ruxolitinib concentrations (Figure 3.7). The 4 hour time point has less variation around the treatment effect than the later time point. The pSTAT5 negative cell lines show a larger degree of variation with some time points showing slight reduction, some slight increase, but no consistent pattern as observed with the positive cell lines. This can be seen in the individual cell lines as well. The two negative cell lines demonstrate a minimal change in expression levels (Figure 3.6). As with the cumulative analysis there is a slight reduction in expression at the lower ruxolitinib concentration, 4 hour time point for some genes, but also some increased expression as well. The five pSTAT5 positive cell lines showed a reduction in expression of most genes at all time-points and concentrations (Figure 3.7). There are a few exceptions where one gene, a different one in individual cell lines, showed an increase in expression. The pattern of transcript down-regulation was present and consistent with both ruxolitinib concentrations and time-points. This finding is consistent with the four signature gene relationship to the activation status of pSTAT5.

The modulation of the 4 genes from signature by ruxolitinib was further examined *in vivo*. To this end, the UKE-1 tumor xenograft model was treated with ruxolitinib. Tumor samples were acquired and analyzed 4 Hr or 24 Hr after single dose of ruxolitinib at 60 mg/kg.

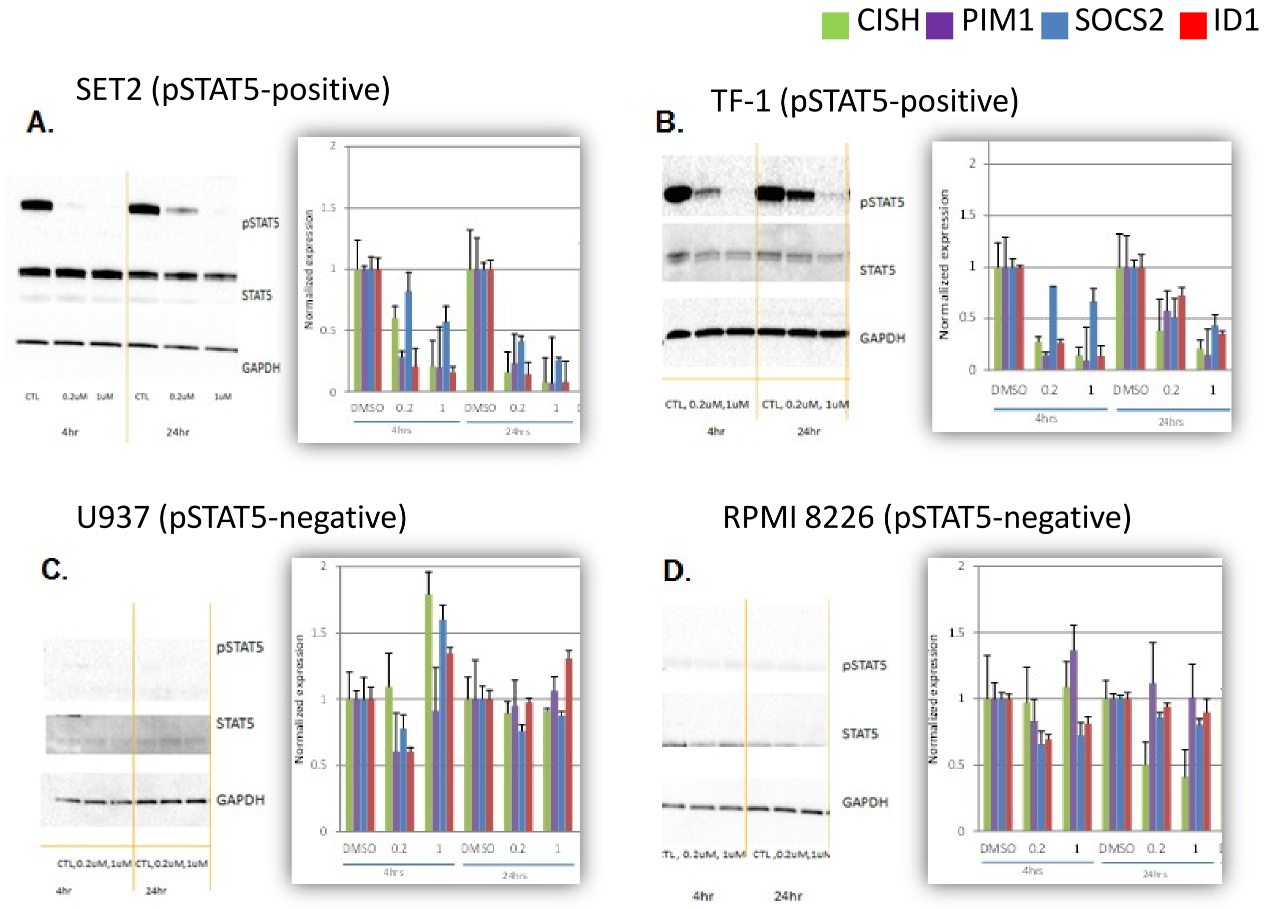


Figure 3.6 Modulation of pSTAT5 and the expression of the signature genes by ruxolitinib

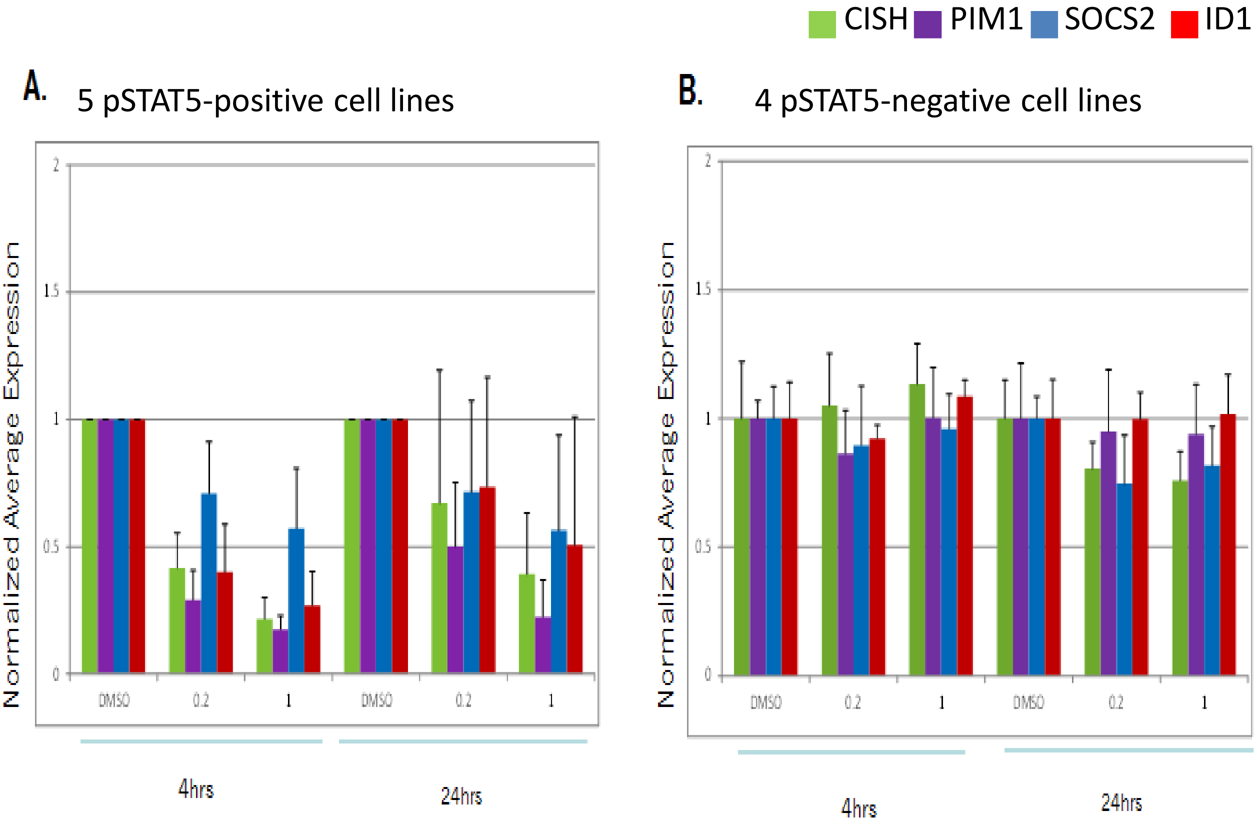


Figure 3.7 Response to ruxolitinib in pSTAT5 positive and negative cell lines

The modulation of pSTAT5 in tumor lysate was examined by Western blotting (Figure 3.8). The modulation of the 4 genes from signature by ruxolitinib in this tumor model is consistent with that observed *in vitro* (Figure 3.8). (The rebound in pSTAT5 level and corresponding rebound in 4 genes mRNA expression is due to decrease in compound concentration.)

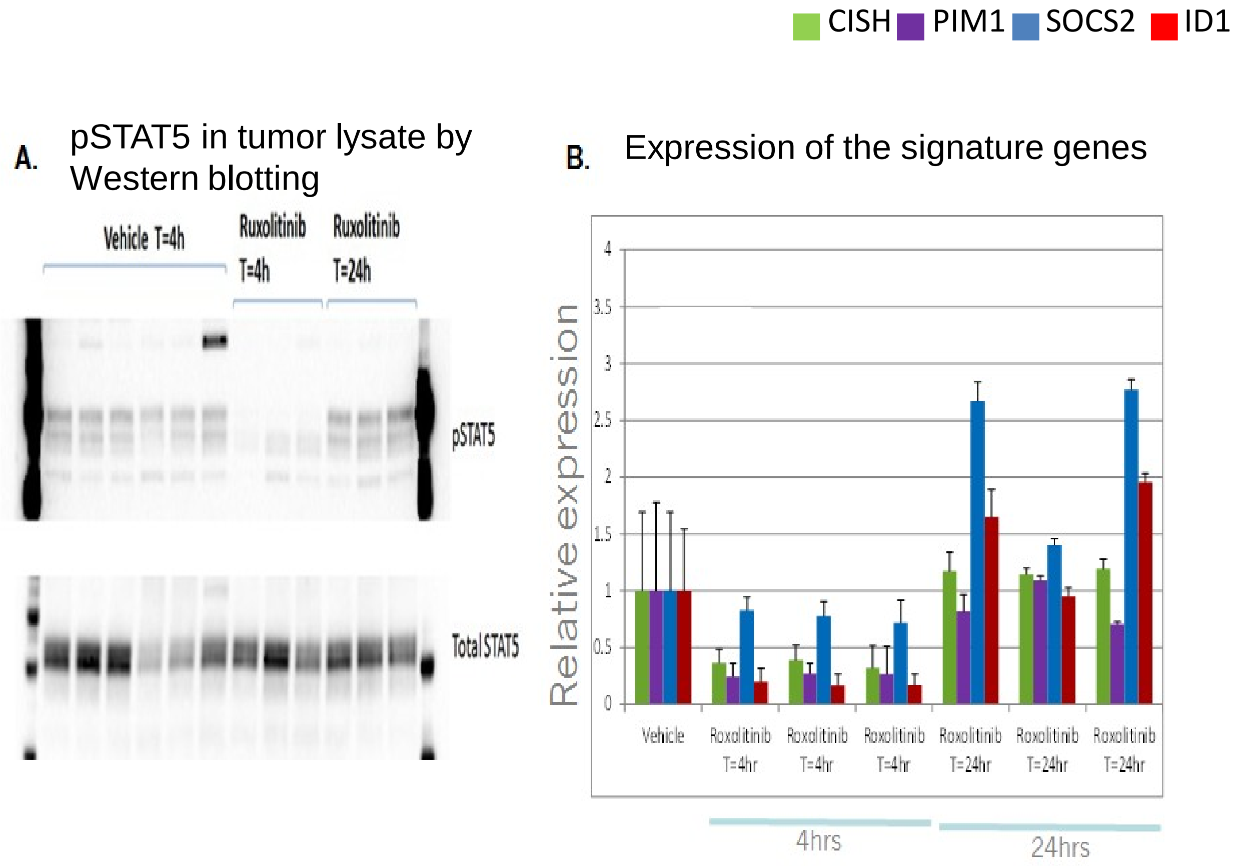


Figure 3.8 Modulation of pSTAT5 by ruxolitinib in the UKE-1 tumor xenograft

3.3.3 BRAF inhibitor sensitivity and MITF transcriptional targets

Another potentially interesting observation based on gene set activity scores pointed to a positive correlation between genes transcriptionally activated by MITF and sensitivity to BRAF inhibitors in V600E mutant melanomas. BRAF is a well-known oncogenic driver and about 50% of melanomas have an activating V600E mutation (Ascierto et al., 2012). PLX-4720 is a selective and potent inhibitor of oncogenic mutant BRAF (Tsai et al., 2008). A high percentage of patients respond to PLX-4720, however unfortunately resistance to the compound often develops in a few months (Flaherty et al., 2010). Most likely there are multiple mechanisms of resistance to the inhibitor. Some mechanisms of resistance are due to mutations reactivating the mitogen-activated protein kinase (MAPK) signaling cascade, for example such as N-RAS mutations (Nazarian et al., 2010) and MEK1 mutations (Wagle et al., 2011). Also the MAPK pathway could be potentially reactivated by MAP3K8 up regulation (Johannessen et al., 2010) and there are instances of resistance to the BRAF inhibitors independent of MAPK pathway (Nazarian et al., 2010). Identifying additional mechanisms of intrinsic resistance to BRAF inhibitors may lead to new therapeutics which would increase the percentage of melanoma patients responding to treatment and also potentially increase the duration of their response.

Table 3.13 shows MITF-Activation gene set scores and PLX-4720 IC50 µM in melanoma cell lines with BRAF V600E mutation. Cell lines with IC50 below 1.5 µM are considered to be sensitive to PLX-4720 while cell lines with IC50 above 4.0 µM are considered to be insensitive to PLX-4720.

Table 3.13 MITF_Activation gene set scores and PLX-4720 IC50 µM

Cell Line Name	MITF_Activation gene set score	PLX-4720 IC50 µM
loximvi	0.64	8.00
wm115	-0.48	8.00
hs294t	2.84	8.00
skmel31	1.2	8.00
rpmi7951	0.12	8.00
wm793	-0.74	8.00
igr39	1.29	8.00
hs695t	0.67	7.26
skmel24	0.14	5.15
colo741	6.7	4.04
mdamb435s	-0.78	2.31
c32	0.37	2.21
k029ax	7.74	2.11
hs939t	3.51	2.10
ht144	1.85	1.34
g361	4.63	1.29
wm1799	2.79	1.23
uacc257	11.7	1.06
igr37	10.19	0.90
rvh421	3.14	0.77
colo679	3.86	0.55
wm983b	1.38	0.51
skmel5	6.18	0.37
melho	10.11	0.31
malme3m	9.01	0.29
a375	0.62	0.26
uacc62	0.56	0.25
wm88	0.25	0.20

Student's t-Test was run using two-tailed distribution and homoscedastic settings for MITF-Activation gene set scores in BRAF inhibitor sensitive cell lines versus BRAF inhibitor insensitive cell lines. The probability associated with the Student's t-Test is 0.02 indicating a potential nonrandom relationship between genes transcriptionally activated by MITF and sensitivity to BRAF inhibitors in V600E mutant melanomas. Interestingly (Tap et al., 2010) also concluded that MITF expression could be a potential predictor of sensitivity to BRAF inhibitors in BRAF mutant melanomas. MITF is a master regulator of melanocyte development and melanoma oncogene (Garraway et al., 2005) (Levy et al., 2006). Appendix A1.6 lists genes transcriptionally activated by MITF.

MITF expression levels are possibly adjusted to satisfy the balance needed to sustain survival and proliferation of melanomas (Goding, 2011). Also it was suggested melanoma cell may have an ability to switch between MITF low and MITF high expression in vivo, with MITF low potentially corresponding to migratory phenotype and MITF high corresponding to proliferative phenotype (Hoek et al., 2008). The underlying plasticity in the MITF expression in melanomas may result in heterogeneous MITF expression in the tumor and potentially be one of the factors in the development of resistance to BRAF inhibitors in BRAF mutant melanomas.

3.4 Discussion

Gene set activity scores on individual samples level represent an interesting approach for analyzing mRNA expression profiles. In the above section such an approach was applied to all cancer cell lines in CCLE collection with available U133Plus2 data.

3.4.1 pSTAT5 gene signature clinical relevance

Gene set activity scores helped to identify potential pSTAT5 gene signature in hematological malignancies. It is important to note that the STAT5 signature is not limited only to aberrant *JAK2* mutations. Other genetic events involving *ABL*, *ALK*, or *FLT3* activation may drive STAT5 activation (Carlesso et al., 1996) (Mizuki et al., 2000) (Zhang et al., 2000) (Hayakawa et al., 2000) (Nieborowska-Skorska et al., 2001), which can be reflected in an elevated signature. For example, K562 is a CML line bearing the Bcr-Abl translocation (Allen et al., 1992). Its STAT5 positivity is likely driven by a high activity of Abl. Also, the fact that myelofibrosis patients with wild type *JAK2* benefited from ruxolitinib treatment support the notion that the JAK-STAT pathway is active in patients with wild type *JAK2* and targeting this pathway is known to yield clinical benefit (Harrison et al., 2012). Furthermore, preclinical findings on possible combination synergy between JAK and PI3K/mTOR inhibitors would suggest that JAK-STAT signaling is a part of the intricate and complex oncogenic network (Maude et al., 2012) (Bogani et al., 2013). The potential utility of pSTAT5 gene signatures to identify subpopulations of patients who may benefit from JAK inhibitor(s) is a rather intriguing possibility.

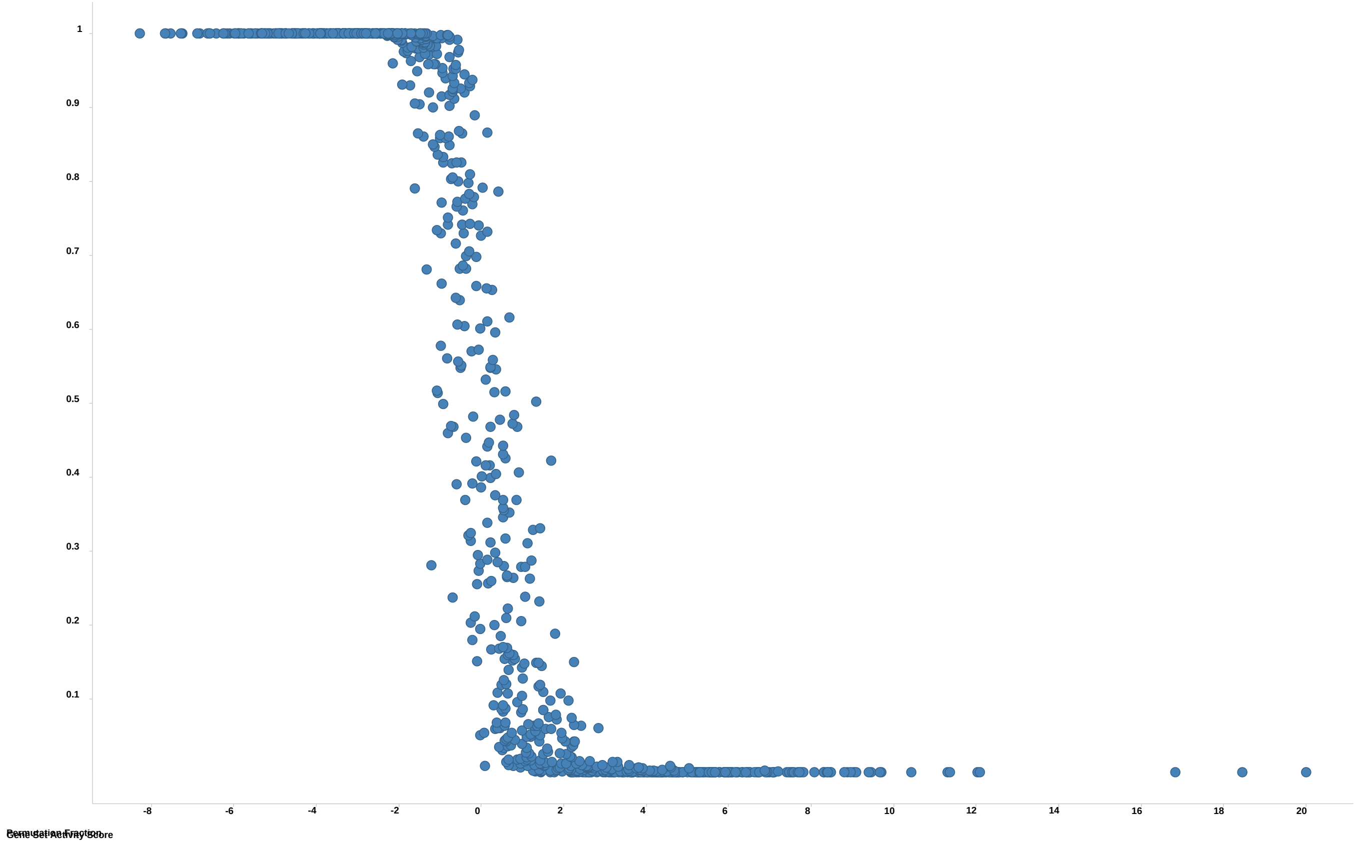


Figure 3.9 Oxidative phosphorylation activity in 964 CCLE cancer cell lines.

3.4.2 Permutation fractions and gene set activity scores distributions

Gene set activity analysis applied to CCLE cancer cell lines mainly took advantage of gene set activity scores, however it is worthwhile to touch on permutation fractions which were also generated for each cell line and gene set.

Oxidative phosphorylation is one of the major mechanisms for energy production in eukaryotic cells and it was previously demonstrated that genes involved in oxidative phosphorylation are tightly co-regulated (van Waveren and Moraes, 2008). Figure 3.9 shows gene set activity scores and permutation fractions for a gene set comprised of genes involved in oxidative phosphorylation for 964 CCLE cancer cell lines. The combination of gene set activity scores and permutation fractions displayed in Figure 3.9 reveals a pattern where the vast majority of cells have co-expression of genes involved in oxidative phosphorylation. The genes are either up regulated together or down regulated together in most cancer cell lines.

Permutation fractions distribution in 964 CCLE lines for 1311 gene sets is shown on Figure 3.10. This figure shows a nearly symmetrical distribution for permutation fractions in relation to the average. The expected number of permutation fractions in each of the 1000 bins could be calculated as (964 x 1311)/1000 which is about 1264 per bin. The bins with more than 3792 permutation fractions could be defined as unusual events, which would approximately correspond to permutation fractions below 0.006 and above 0.997. Permutation fractions distribution density is more concentrated at the right tail of distribution in comparison to the left tail of distribution.

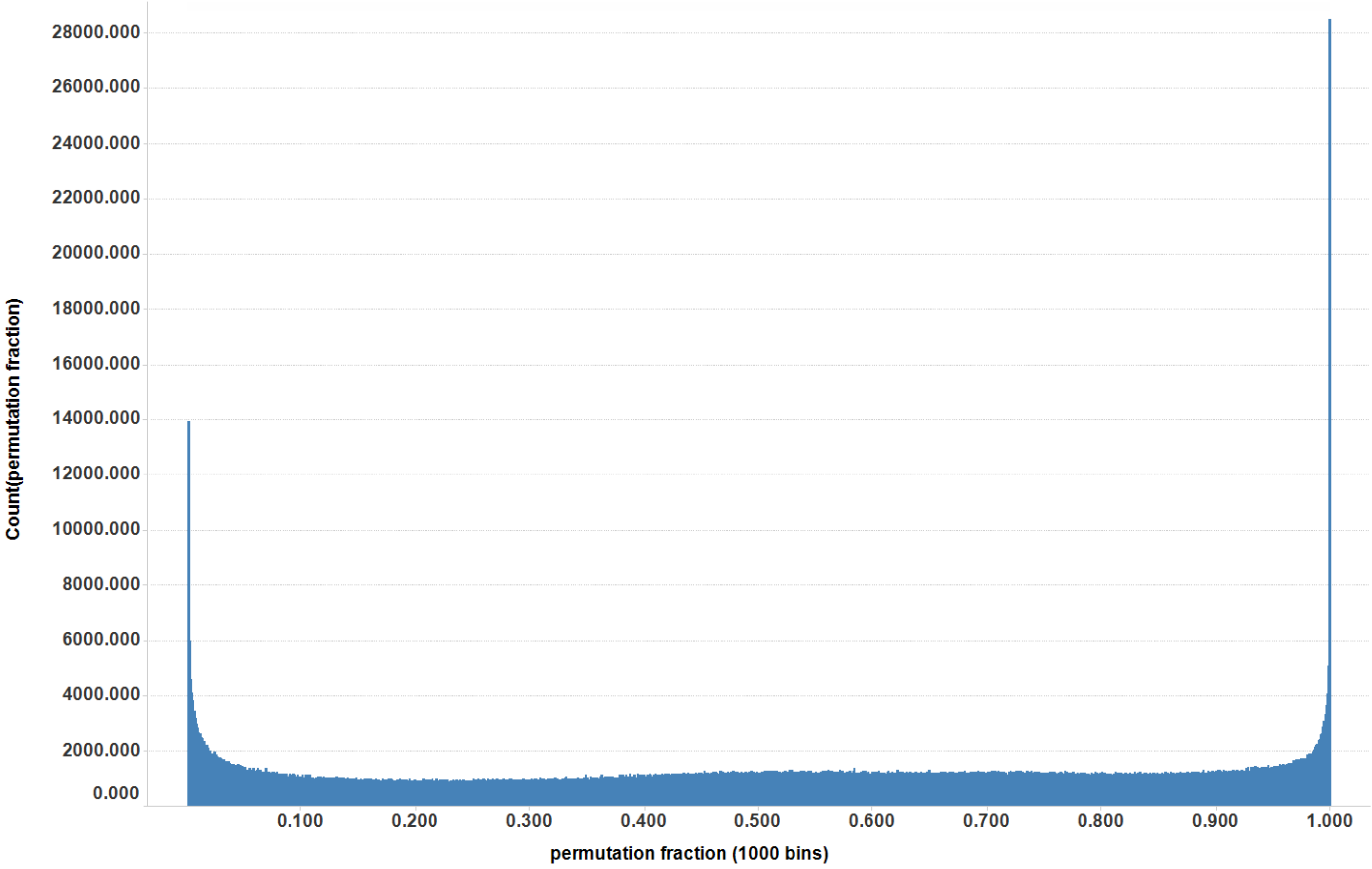


Figure 3.10 Permutation fractions distribution in 964 CCLE lines for 1311 gene sets.

Figure 3.11 shows gene set activity scores distribution in 964 CCLE lines for 1311 gene sets. As expected the mean of the activity scores distribution is around zero and standard deviation is around one. Figure 3.12 shows gene set activity scores distribution in -5.00 to 5.00 range, in that window which covers vast majority of data points the distribution is nearly symmetrical.

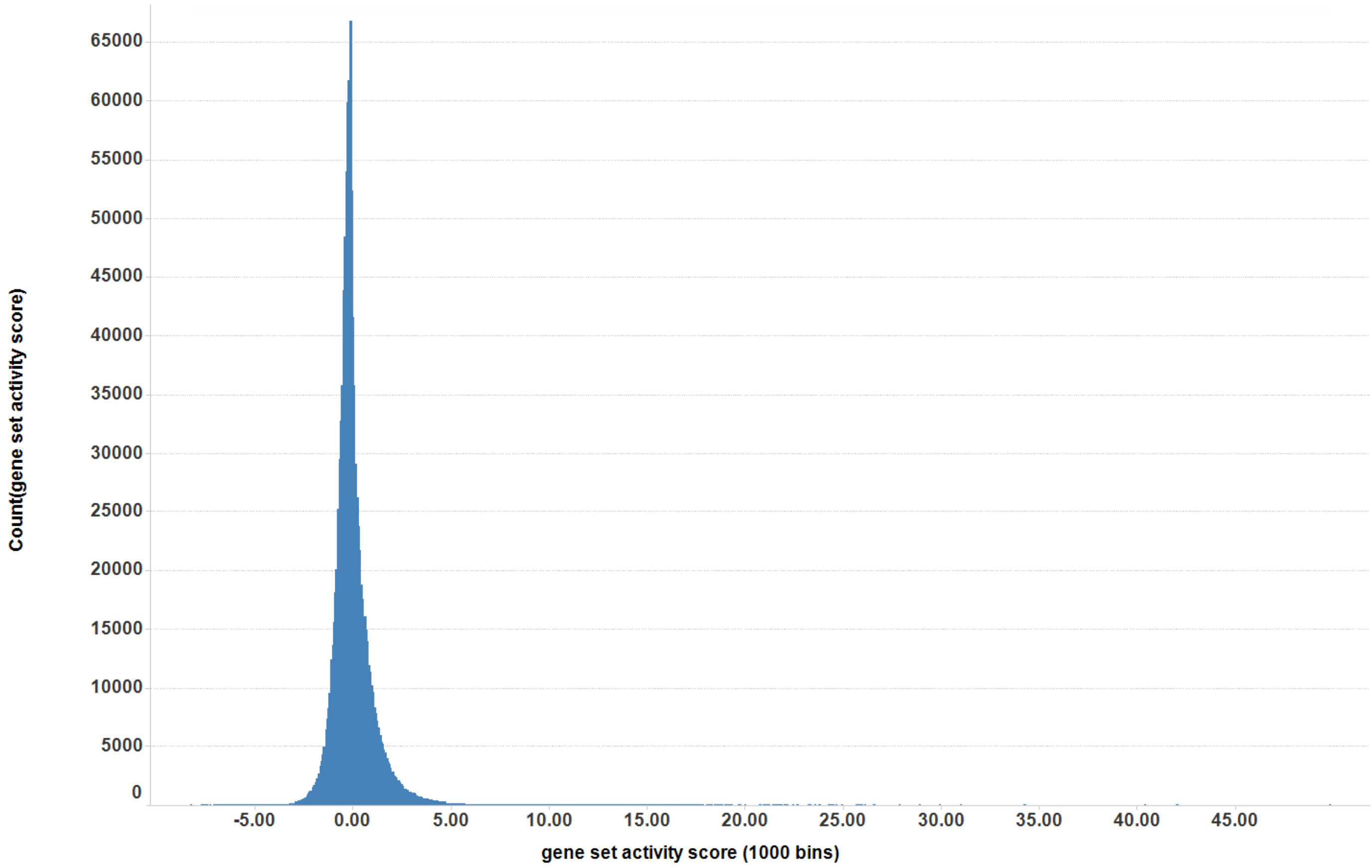


Figure 3.11 Gene set activity scores distribution in 964 CCLE lines for 1311 gene sets.

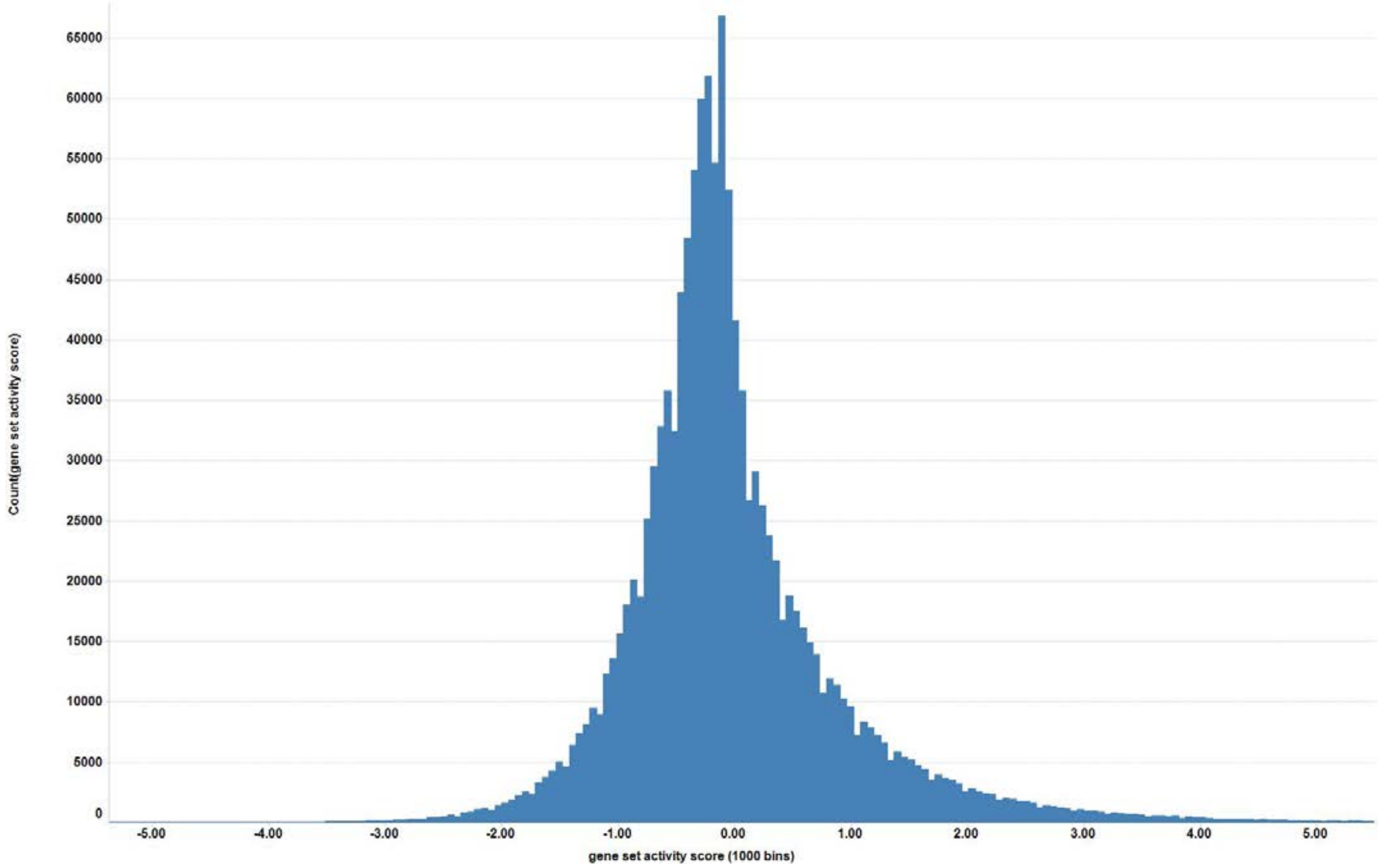


Figure 3.12 Gene set activity scores distribution in -5.00 to 5.00 range.
(964 CCLE lines for 1311 gene sets)

Permutation fractions relation to gene set activity scores in 964 CCLE lines for 1311 gene sets is shown on Figure 3.13. This figure shows overall reasonable concordance between permutation fractions and gene set activity scores.

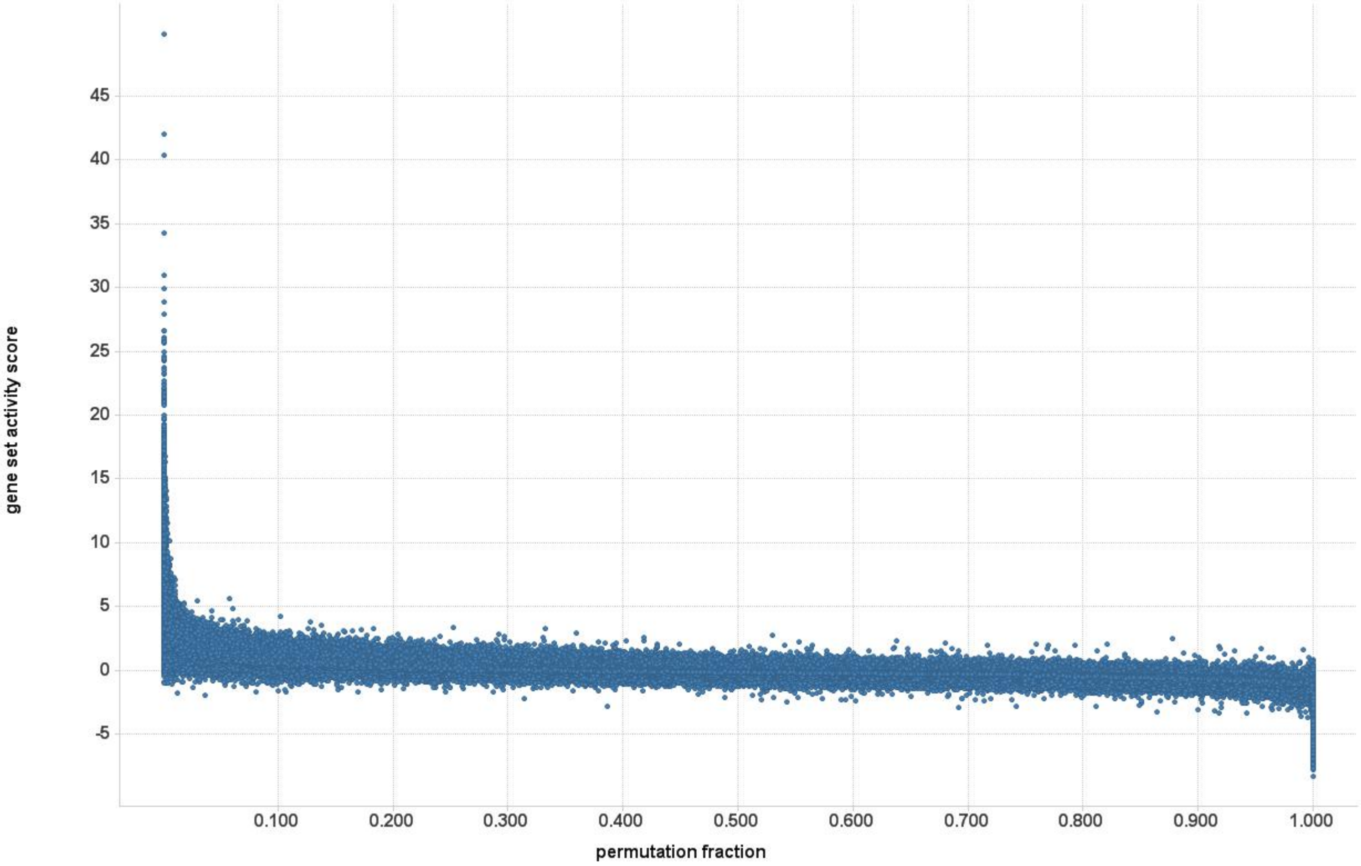


Figure 3.13 Permutation fractions versus gene set activity scores in 964 CCLE lines.

3.4.3 Z-scores based FDR-corrected p-values

Since Figure 3.12 shows that distribution of gene set activity scores for 1311 gene sets across 964 cell lines have a shape resembling a normal distribution, we can therefore try to estimate gene set activity score threshold which would correspond to p-value of 0.05 under False Discovery Rate (FDR) control. Under FDR control for 1311 gene sets the nominal p-value would be 0.05/1311 which approximately equal to 3.8E-05 and corresponding z-score for two-tailed test is approximately equal to 4.12. As can be seen from Figure 3.13 vast majority of gene set activity scores above 4.12 have permutation fraction < 0.006 and also vast majority of gene set activity scores below -4.12 have permutation fraction > 0.997. Therefore there is a reasonable concordance between calculations based on z-scores and calculations based on permutation fractions. Appendix A2 CD contains a file with nominal p-values based on z-scores and file with corresponding FDR corrected p-values (Hochberg and Benjamini, 1990) for 1311 gene sets across all 964 CCLE cell lines. Figure 3.14 shows permutation fractions distribution for corresponding z-scores based FDR-corrected p-values below 0.05 for 1311 gene sets across all 964 CCLE cell lines. This figure indicates that concordance between permutation fractions and corresponding z-scores based FDR-corrected p-values below 0.05 is better at the left tail of distribution in comparison to the right tail of distribution.

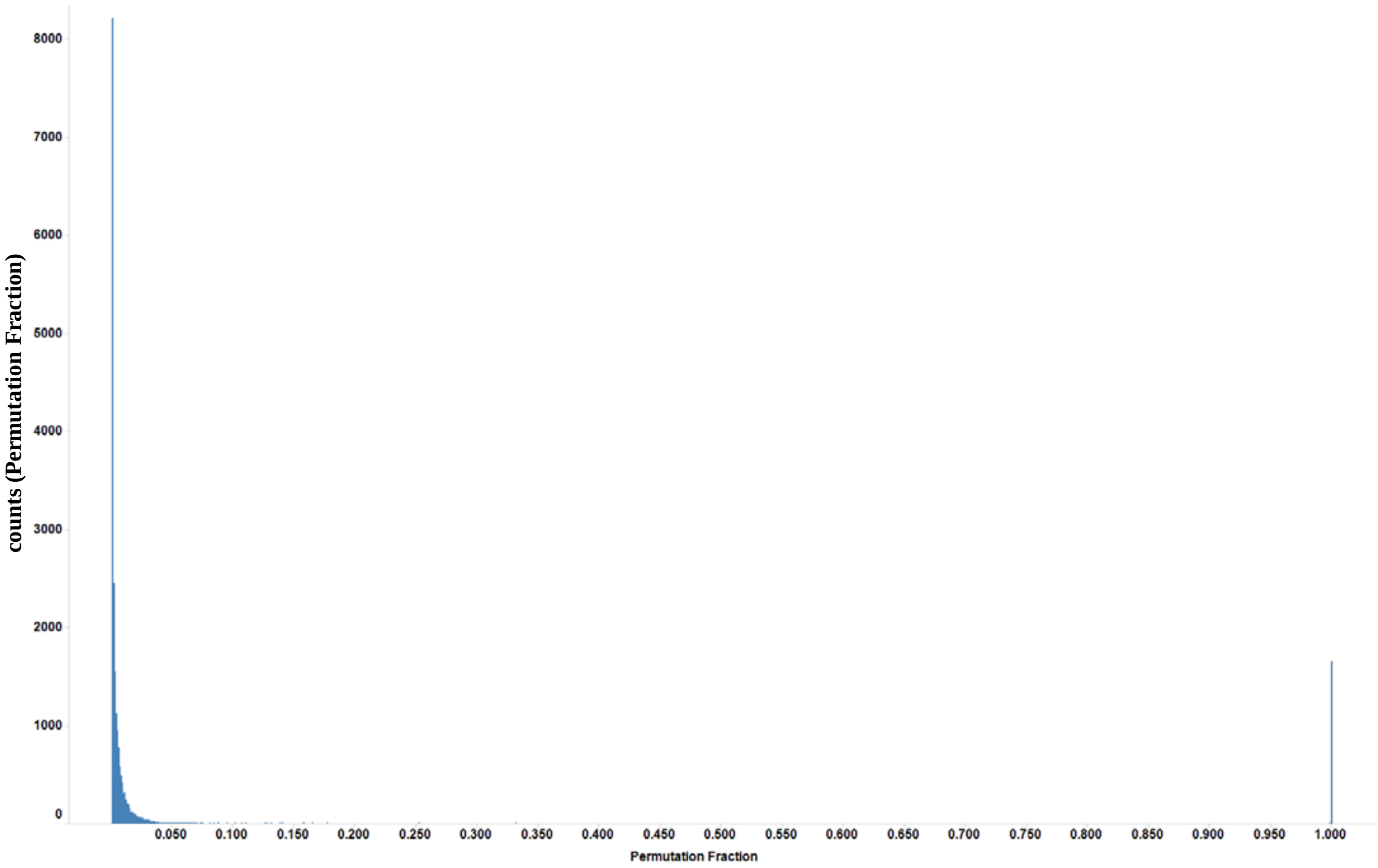


Figure 3.14 Permutation fractions distribution for z-scores based FDR-corrected p-values < 0.05.

There are number of challenges in generating and interpreting gene set activity scores. Unusually high or low activity scores allow us to focus our attention on potentially more relevant biological processes. However it does not necessarily mean that scores which do not stand out are not providing important information. For example it is possible to imagine a scenario where a particular pathway has mediocre activity scores and values between different samples that do not vary more than 1.5 times; yet such changes could still have far reaching effects on the cell. Also there are number of the gene sets defined based on signaling pathways which depend on posttranslational modifications and which in turn may not be reflected correctly or at all on the mRNA expression level. Activity scores for directional transcriptional gene sets often provide the most relevant read outs of biological activity; however they also can be diluted by transcriptional targets on which they have little effect or on the contrary some important transcriptional targets may not yet be known or not have been accounted for. Other possible complications are the differences between different lineages, such differences may mean that some gene sets need to have different composition for different lineages. On another hand the challenges outlined above could also potentially indicate that future improvements in gene set compositions and interpretation may increase the relevance of gene set activity analysis.

3.5 Summary

This chapter described details of the computational framework for the gene set activity analysis on sample by sample basis. The framework has been developed as part of doctoral research. It was applied to mRNA expression data analysis from Cancer Cell Line Encyclopedia. Tissue specific genes and tissue specific processes were used to validate analysis results. Generation of pSTAT5 mRNA expression signature demonstrated utility of gene set activity analysis on sample by sample basis. Potential clinical relevance of pSTAT5 gene signature was highlighted. An interesting relationship between BRAF inhibitor sensitivity and MITF signaling was discovered by taking advantage of gene set activity analysis on sample by sample basis.

Chapter 4: General Discussion

During the last few decades significant progress has been made in understanding molecular interactions involved in various cancers. In recent years due to the exponential increase in DNA sequencing capacity and corresponding exponential decrease in cost of DNA sequencing, the amount of available genomic information for oncology research has greatly increased. Also, for the first time significant numbers of samples with multiple genomic data sets per sample have became available. These introduce the challenge and at the same time the opportunity for improving existing computational approaches in analyzing these new and potentially information-rich data sets.

There are number of a ways how availability of multiple genomic data sets per sample can be utilized. For example, gene fusion detection could be done using RNA-seq data and whole genome sequencing data. Whole genome sequencing data allows detection of a wide range of fusions and quantifies amplification of potential fusion products, but it does not provide information on actual expression of fusion products. RNA-seq data on other hand provides mRNA expression of fusion product and also can detect read through fusions (Akiva et al., 2006) which cannot be detected from whole genome sequencing data. At the same time RNA-seq data may introduce a number of false positive results due to complex isoform structures and other technical and biological reasons (Asmann et al., 2011). Combining fusion detection from RNA-seq data and whole genome sequencing data helps to decrease the number of false positive events and also to narrow down the search to the most relevant fusions (McPherson et al., 2011).

4.1 Tumor suppressor genes status analysis

The comprehensive and systematic computational framework to interrogate tumor suppressor genes status presented in this thesis contributes to improvement in understanding of cancer cell lines which play critical role in cancer research. The comprehensive nomenclature introduced in this work allows capturing in one short string the underlying status of tumor suppressor for gene and sample in question. It is important to note that prior to this work there were no published comprehensive nomenclature system for tumor suppressor status. The tumor suppressor status using this systematic nomenclature could be used as input in other computational data analysis methods. For example, tumor suppressor genes status could be used as one of the inputs in machine learning algorithms trying to predict sensitivity to number of different anticancer compounds. In fact status of tumor suppressor genes generated for CCLE cancer cell lines are already extensively used in variety of approaches for target identification and biomarker discovery.

Improvements in characterizing tumor suppressor status, made possible due to work presented in this thesis, are relevant for selecting with high levels of confidence cell lines with functional (wild type) tumor suppressor genes and also on another side of the spectra selecting with high confidence cell lines with inactivated tumor suppressor genes. Identifying cell lines with wild type TP53 is critical for efforts to identify signature to predict sensitivity to inhibitors of MDM2 driven degradation of TP53. On the other hand identifying cell lines with inactivated BRCA1 or BRCA2 are important for efforts to widen therapeutic benefits from PARP inhibitors. The status of 69 tumor suppressor genes across almost 800 CCLE cancer cell lines have already been shared with the scientific community through publication in the *Journal of Molecular Oncology* providing a valuable resource for years to come (Sonkin et al., 2013).

The work on tumor suppressor status framework also adds to the number of voices advocating the creation of a comprehensive and reliable resource for annotation of loss of function as well as gain of function missense mutations in cancer. Such a resource would be very useful for the field of oncology research, especially keeping in mind the increasing volume of sequencing data available for basic, translational and clinical research.

In the future several additional genomic data sets are going to be available for cancer cell lines, most notably the RNA-seq data and DNA methylation database. Incorporating this and other additional data sets would allow further improvements in interrogation of tumor suppressor status. The RNA-seq data would allow detecting potential allele specific expression and therefore provide more precise tumor suppressor status in cases than wild type and mutant alleles of genes in question are present. It also may allow detection of novel isoforms which lead to loss of wild type functionality. DNA methylation data on the other hand may provide potential mechanistic explanation and additional layer of evidence to loss of expression of gene in question in case of absence of any other detectable genetic alterations. Whole genome sequencing data at sufficient depth is eventually going to be available for cancer cell lines, and combining it with all other genomic data sets should provide comprehensive genetic characterization of cell lines. The detailed analysis of promoter and enhancer regions may provide evidence on additional mechanisms of tumor suppressor genes inactivation. Such characterization may also detect more instances of cancer with tumor suppressor genes inactivation and no oncogene alterations. In depth investigation of such cases may help to better understand underlying signaling pathways and highlight potential opportunities for therapeutic interventions.

The same approach introduced here for the analysis of tumor suppressor status in cancer cell lines can, with some modifications, be extended to analysis of primary tumor samples and xenografts established from primary tumor samples. Such analysis should allow more precise characterization of tumor suppressor status in samples in question.

The framework introduced in this work and outlined in Figure 2.3 was designed specifically to help better define status of tumor suppressor genes in cancer. However the framework could be adapted for applications outside of the oncology field. For example some of the genetic syndromes are caused by loss of function of variety of genes, and genes could be inactivated by different mechanisms and to different degrees (Strachan et al., 2011) (Ségalat, 2007). In fact the amount of potential remaining functionality of the gene or genes in question in number of cases is related to severity of symptoms (Nussbaum et al., 2007). With appropriate modification the existing framework can help to account for multiple mechanisms of genes inactivation and also potentially account for degree of inactivation in genetic syndromes. Also since some of the phenotypes associated with genetic syndromes could be caused by loss of function of one of the genes from the set of multiple genes, it is possible to consider generating aggregative status call covering the whole set of genes.

4.2 Gene sets activity analysis on sample by sample basis

Gene set activity scores calculated in this thesis for nearly one thousand cancer cell lines, based on mRNA expression profiles, allow another angle in characterization of transcriptional read outs. It is important to note that prior to this work generating sample based gene set activity scores for such extensive collection of cancer cell lines have not been previously attempted. This work demonstrated that it is possible to perform sample based gene set activity analysis across thousands of samples. It also added additional support to the notion that z-score based gene set activity analysis is a solid and practical approach. The framework presented here for assessing gene set activity scores on sample by sample basis is regularly used to investigate epidemiology of variety of pathways across hundreds of cell lines and thousands of primary and metastatic tumor samples.

This framework is a good platform for investigating various potential future improvements. One of the key areas for future improvements is the enhancement of membership of gene sets. In general it is easier to interpret gene set activity scores than genes in the gene set for gene set activity analysis based on mRNA data are transcriptionally regulated in the same direction by the biological process in question. In order to achieve such memberships for canonical signaling pathways the computational approach needs to be developed to distil signaling pathway in question to its core of transcriptionally regulated genes. Such work would likely need to be followed up by wet lab experiments in order to further confirm and refine gene set memberships.

Gene set activity scores were used in part to define pSTAT5 mRNA expression signature in hematopoietic cancer cell lines. This signature can potentially make it possible to identify patients whom may benefit from JAK inhibitor(s), based on JAK-STAT signaling. The poster outlining the pSTAT5 mRNA expression signature have been presented at 2013 American Society of Clinical Oncology (ASCO) annual meeting and it was downloaded more than 200 times just in a span of few days, indicating the relevance of work to basic, translational and clinical oncology research.

4.3 Summary

Two main research projects formed a foundation of the presented thesis: gene set activity analysis on sample by sample basis and a tumor suppressor genes status analysis.

Computational framework developed for gene set activity analysis allows integrating knowledge of signaling pathways with mRNA expression profiles available for each sample of extensive collection of cancer cell lines from Cancer Cell Line Encyclopedia. This framework was utilized for generating mRNA expression based signature reflecting phosphorylation status of STAT5. Antibody based approach to access pSTAT5 status is complicated by multiple technical and logistic challenges which greatly decreases availability of such analysis in clinical practice. Expression-based assessment of pSTAT5 status promises potentially higher availability of such analysis in clinic. Knowledge of pSTAT5 status based on mRNA expression signature has a potential to help to select patients for treatment by JAK2 inhibitors.

Comprehensive and systematic computational framework developed for tumor suppressor genes status analysis takes the integration theme a few steps further. The framework takes advantage of mutation, copy number and mRNA expression data available for cancer cell lines in Cancer Cell Line Encyclopedia. Availability of multiple data sets per individual cell line allows the framework to account for multiple mechanisms of tumor suppressor genes inactivation. Accounting for multiple mechanisms of inactivation of tumor suppressor genes improves genomic characterization of cancer cell lines which are important and heavily used models for cancer research. In particular the improvement in characterization of TP53 status in cancer cell lines due to work presented in this thesis creates a better foundation for research to predict which patients would benefit from treatment with inhibitors of MDM2 driven TP53 degradation.

References

Agirre, X., Novo, F.J., Calasanz, M.J., Larráyoz, M.J., Lahortiga, I., Valgañón, M., García-Delgado, M., Vizmanos, J.L., 2003. TP53 is frequently altered by methylation, mutation, and/or deletion in acute lymphoblastic leukaemia. Mol. Carcinog. 38, 201–208.

Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A., Sorek, R., 2006. Transcription-mediated gene fusion in the human genome. Genome Res. 16, 30–36.

Allen, P.B., Morgan, G.J., Wiedemann, L.M., 1992. Philadelphia chromosome-positive leukaemia: the translocated genes and their gene products. Baillieres Clin. Haematol. 5, 897–930.

Al-Shahrour, F., Díaz-Uriarte, R., Dopazo, J., 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20, 578–580.

Amaravadi, R., Thompson, C.B., 2005. The survival kinases Akt and Pim as potential pharmacological targets. J. Clin. Invest. 115, 2618–2624.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J., 2002. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. 30, 41–47.

Ascierto, P.A., Kirkwood, J.M., Grob, J.-J., Simeone, E., Grimaldi, A.M., Maio, M., Palmieri, G., Testori, A., Marincola, F.M., Mozzillo, N., 2012. The role of BRAF V600 mutation in melanoma. J Transl Med 10, 85.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Asmann, Y.W., Hossain, A., Necela, B.M., Middha, S., Kalari, K.R., Sun, Z., Chai, H.-S., Williamson, D.W., Radisky, D., Schroth, G.P., Kocher, J.-P.A., Perez, E.A., Thompson, E.A., 2011. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res. 39, e100.

Bachmann, M., Möröy, T., 2005. The serine/threonine kinase Pim-1. Int. J. Biochem. Cell Biol. 37, 726–730.

Bale, S., Devisscher, M., Van Criekinge, W., Rehm, H.L., Decouttere, F., Nussbaum, R., Dunnen, J.T.D., Willems, P., 2011. MutaDATABASE: a centralized and standardized DNA variation database. Nat. Biotechnol. 29, 117–118.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E.M., Sos, M.L., Michel, K., Mermel, C., Silver, S.J., Weir, B.A., Reiling, J.H., Sheng, Q., Gupta, P.B., Wadlow, R.C., Le, H., Hoersch, S., Wittner, B.S., Ramaswamy, S., Livingston, D.M., Sabatini, D.M., Meyerson, M., Thomas, R.K., Lander, E.S., Mesirov, J.P., Root, D.E., Gilliland, D.G., Jacks, T., Hahn, W.C., 2009. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462, 108–112.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., et al., 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607.

Bauer-Mehren, A., Furlong, L.I., Sanz, F., 2009. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol. Syst. Biol. 5, 290.

Baxter, E.J., Scott, L.M., Campbell, P.J., East, C., Fourouclas, N., Swanton, S., Vassiliou, G.S., Bench, A.J., Boyd, E.M., Curtin, N., Scott, M.A., Erber, W.N., Green, A.R., Cancer Genome Project, 2005. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. Lancet 365, 1054–1061.

Beltrame, L., Bianco, L., Fontana, P., Cavalieri, D., 2013. Pathway-based analysis of microarray and RNAseq data using Pathway Processor 2.0. Curr Protoc Bioinformatics Chapter 7, Unit 7.6.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., et al., 2010. The landscape of somatic copy-number alteration across human cancers. Nature 463, 899–905.

Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J.M., Berchuck, A., Olson, J.A., Jr, Marks, J.R., Dressman, H.K., West, M., Nevins, J.R., 2006. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439, 353–357.

Bogani, C., Bartalucci, N., Martinelli, S., Tozzi, L., Guglielmelli, P., Bosi, A., Vannucchi, A.M., Associazione Italiana per la Ricerca sul Cancro AGIMM Gruppo Italiano Malattie Mieloproliferative, 2013. mTOR inhibitors alone and in combination with JAK2 inhibitors effectively inhibit cells of myeloproliferative neoplasms. PLoS ONE 8, e54826.

Breslin, T., Krogh, M., Peterson, C., Troein, C., 2005. Signal transduction pathway profiling of individual tumor samples. BMC Bioinformatics 6, 163.

Brown, C.J., Lain, S., Verma, C.S., Fersht, A.R., Lane, D.P., 2009. Awakening guardian angels: drugging the p53 pathway. Nat. Rev. Cancer 9, 862–873.

Cancer Genome Atlas Research Network, 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068.

Cancer Genome Atlas Research Network, 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. 368, 2059–2074.

Carbone, M., Ferris, L.K., Baumann, F., Napolitano, A., Lum, C.A., Flores, E.G., Gaudino, G., Powers, A., Bryant-Greenwood, P., Krausz, T., Hyjek, E., Tate, R., Friedberg, J., Weigel, T., Pass, H.I., Yang, H., 2012. BAP1 cancer syndrome: malignant mesothelioma, uveal and cutaneous melanoma, and MBAITs. J Transl Med 10, 179.

Carlesso, N., Frank, D.A., Griffin, J.D., 1996. Tyrosyl phosphorylation and DNA binding activity of signal transducers and activators of transcription (STAT) proteins in hematopoietic cell lines transformed by Bcr/Abl. J. Exp. Med. 183, 811–820.

Cheadle, C., Vawter, M.P., Freed, W.J., Becker, K.G., 2003. Analysis of microarray data using Z score transformation. J Mol Diagn 5, 73–81.

Chellappan, S.P., Hiebert, S., Mudryj, M., Horowitz, J.M., Nevins, J.R., 1991. The E2F transcription factor is a cellular target for the RB protein. Cell 65, 1053–1061.

Chen, F., Wang, W., El-Deiry, W.S., 2010. Current strategies to target p53 in cancer. Biochem. Pharmacol. 80, 724–730.

Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M., Halfon, M.S., 2005. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol. 6, R16.

Cortes, J.E., Kantarjian, H., Shah, N.P., Bixby, D., Mauro, M.J., Flinn, I., O'Hare, T., Hu, S., Narasimhan, N.I., Rivera, V.M., Clackson, T., Turner, C.D., Haluska, F.G., Druker, B.J., Deininger, M.W.N., Talpaz, M., 2012. Ponatinib in refractory Philadelphia chromosome-positive leukemias. N. Engl. J. Med. 367, 2075–2088.

Croce, C.M., 2008. Oncogenes and cancer. N. Engl. J. Med. 358, 502–511.

Dangles-Marie, V., Pocard, M., Richon, S., Weiswald, L.-B., Assayag, F., Saulnier, P., Judde, J.-G., Janneau, J.-L., Auger, N., Validire, P., Dutrillaux, B., Praz, F., Bellet, D., Poupon, M.-F., 2007. Establishment of human colon cancer cell lines from fresh tumors versus xenografts: comparison of success rate and cell line features. Cancer Res. 67, 398–407.

Dennis, G., Jr, Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4, P3.

Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulski, K.S., Halloran, P., Yasui, Y., 2007. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics 8, 242.

Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C., Conklin, B.R., 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. 4, R7.

Druker, B.J., Talpaz, M., Resta, D.J., Peng, B., Buchdunger, E., Ford, J.M., Lydon, N.B., Kantarjian, H., Capdeville, R., Ohno-Jones, S., Sawyers, C.L., 2001. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N. Engl. J. Med. 344, 1031–1037.

Dudoit, S., Shaffer, J.P., Block, J.C., 2003. Multiple Hypothesis Testing in Microarray Experiments. Statistical Science 18, 71–103.

Edelman, E., Porrello, A., Guinney, J., Balakumaran, B., Bild, A., Febbo, P.G., Mukherjee, S., 2006. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. Bioinformatics 22, e108–116.

Efeyan, A., Ortega-Molina, A., Velasco-Miguel, S., Herranz, D., Vassilev, L.T., Serrano, M., 2007. Induction of p53-dependent senescence by the MDM2 antagonist nutlin-3a in mouse cells of fibroblast origin. Cancer Res. 67, 7350–7357.

Endo, T.A., Masuhara, M., Yokouchi, M., Suzuki, R., Sakamoto, H., Mitsui, K., Matsumoto, A., Tanimura, S., Ohtsubo, M., Misawa, H., Miyazaki, T., Leonor, N., Taniguchi, T., Fujita, T., Kanakura, Y., Komiya, S., Yoshimura, A., 1997. A new protein containing an SH2 domain that inhibits JAK kinases. Nature 387, 921–924.

Fang, F., Antico, G., Zheng, J., Clevenger, C.V., 2008. Quantification of PRL/Stat5 signaling with a novel pGL4-CISH reporter. BMC Biotechnol. 8, 11.

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N.J., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., Martin, N.M.B., Jackson, S.P., Smith, G.C.M., Ashworth, A., 2005. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 434, 917–921.

Finn, R.S., Dering, J., Conklin, D., Kalous, O., Cohen, D.J., Desai, A.J., Ginther, C., Atefi, M., Chen, I., Fowst, C., Los, G., Slamon, D.J., 2009. PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. Breast Cancer Research 11, R77.

Flaherty, K.T., Puzanov, I., Kim, K.B., Ribas, A., McArthur, G.A., Sosman, J.A., O'Dwyer, P.J., Lee, R.J., Grippo, J.F., Nolop, K., Chapman, P.B., 2010. Inhibition of mutated, activated BRAF in metastatic melanoma. N. Engl. J. Med. 363, 809–819.

Fodde, R., 2002. The APC gene in colorectal cancer. Eur. J. Cancer 38, 867–871.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J.W., Campbell, P.J., Stratton, M.R., Futreal, P.A., 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 39, D945–950.

Freedman, D., 2005. Statistical models: theory and practice. Cambridge University Press, Cambridge; New York.

Friend, S.H., Bernards, R., Rogelj, S., Weinberg, R.A., Rapaport, J.M., Albert, D.M., Dryja, T.P., 1986. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature 323, 643–646.

Gambacorti-Passerini, C., Antolini, L., Mahon, F.-X., Guilhot, F., Deininger, M., Fava, C., Nagler, A., Della Casa, C.M., Morra, E., Abruzzese, E., D'Emilio, A., Stagno, F., le Coutre, P., Hurtado-Monroy, R., Santini, V., Martino, B., Pane, F., Piccin, A., Giraldo, P., Assouline, S., Durosinmi, M.A., Leeksma, O., Pogliani, E.M., Puttini, M., Jang, E., Reiffers, J., Valsecchi, M.G., Kim, D.-W., 2011. Multicenter independent assessment of outcomes in chronic myeloid leukemia patients treated with imatinib. J. Natl. Cancer Inst. 103, 553–561.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., et al., 2012. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483, 570–575.

Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J., Lee, C., Wagner, S.N., Li, C., Golub, T.R., Rimm, D.L., Meyerson, M.L., Fisher, D.E., Sellers, W.R., 2005. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 436, 117–122.

Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S.M., Aburatani, H., 2005. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics 86, 127–141.

Gey, G.O., Coffmann, W.D., Kubicek, M.T., 1952. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. Cancer Res. 12, 264–265.

Gibbons, J.D., 2003. Nonparametric statistical inference. Marcel Dekker, New York.

Gillet, J.-P., Varma, S., Gottesman, M.M., 2013. The clinical relevance of cancer cell lines. J. Natl. Cancer Inst. 105, 452–458.

Goding, C.R., 2011. Commentary. A picture of Mitf in melanoma immortality. Oncogene 30, 2304–2306.

Goeman, J.J., van de Geer, S.A., de Kort, F., van Houwelingen, H.C., 2004. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20, 93–99.

Goodrich, D.W., 2006. The retinoblastoma tumor-suppressor gene, the exception that proves the rule. Oncogene 25, 5233–5243.

Gozgit, J.M., Bebernitz, G., Patil, P., Ye, M., Parmentier, J., Wu, J., Su, N., Wang, T., Ioannidis, S., Davies, A., Huszar, D., Zinda, M., 2008. Effects of the JAK2 inhibitor, AZ960, on Pim/BAD/BCL-xL survival signaling in the human JAK2 V617F cell line SET-2. J. Biol. Chem. 283, 32334–32343.

Groffen, J., Stephenson, J.R., Heisterkamp, N., de Klein, A., Bartram, C.R., Grosveld, G., 1984. Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. Cell 36, 93–99.

Grosu, P., Townsend, J.P., Hartl, D.L., Cavalieri, D., 2002. Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. Genome Res. 12, 1121–1126.

Guo, Z., Zhang, T., Li, X., Wang, Q., Xu, J., Yu, H., Zhu, J., Wang, H., Wang, C., Topol, E.J., Wang, Q., Rao, S., 2005. Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinformatics 6, 58.

Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Kronnie, G.T., Béné, M.-C., De Vos, J., Hernández, J.M., Hofmann, W.-K., Mills, K.I., Gilkes, A., Chiaretti, S., Shurtleff, S.A., Kipps, T.J., Rassenti, L.Z., Yeoh, A.E., Papenhausen, P.R., Liu, W.-M., Williams, P.M., Foà, R., 2010. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group. J. Clin. Oncol. 28, 2529–2537.

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

Hänzelmann, S., Castelo, R., Guinney, J., 2013. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 14, 7.

Harrison, C., Kiladjian, J.-J., Al-Ali, H.K., Gisslinger, H., Waltzman, R., Stalbovskaya, V., McQuitty, M., Hunter, D.S., Levy, R., Knoops, L., Cervantes, F., Vannucchi, A.M., Barbui, T., Barosi, G., 2012. JAK inhibition with ruxolitinib versus best available therapy for myelofibrosis. N. Engl. J. Med. 366, 787–798.

Hayakawa, F., Towatari, M., Kiyoi, H., Tanimoto, M., Kitamura, T., Saito, H., Naoe, T., 2000. Tandem-duplicated Flt3 constitutively activates STAT5 and MAP kinase and introduces autonomous cell growth in IL-3-dependent cell lines. Oncogene 19, 624–631.

Heard, E., Clerc, P., Avner, P., 1997. X-chromosome inactivation in mammals. Annu. Rev. Genet. 31, 571–610.

Hedvat, M., Huszar, D., Herrmann, A., Gozgit, J.M., Schroeder, A., Sheehy, A., Buettner, R., Proia, D., Kowolik, C.M., Xin, H., Armstrong, B., Bebernitz, G., Weng, S., Wang, L., Ye, M., McEachern, K., Chen, H., Morosini, D., Bell, K., Alimzhanov, M., Ioannidis, S., McCoon, P., Cao, Z.A., Yu, H., Jove, R., Zinda, M., 2009. The JAK2 inhibitor AZD1480 potently blocks Stat3 signaling and oncogenesis in solid tumors. Cancer Cell 16, 487–497.

Hochberg, Y., Benjamini, Y., 1990. More powerful procedures for multiple significance testing. Stat Med 9, 811–818.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., McCombie, W.R., 2007. Genome-wide in situ exon capture for selective resequencing. Nat. Genet. 39, 1522–1527.

Hoek, K.S., Eichhoff, O.M., Schlegel, N.C., Döbbeling, U., Kobert, N., Schaerer, L., Hemmi, S., Dummer, R., 2008. In vivo switching of human melanoma cells between proliferative and invasive states. Cancer Res. 68, 650–656.

Hollstein, M., Hainaut, P., 2010. Massively regulated genes: the example of TP53. J. Pathol. 220, 164–173.

Huang, P.-I., Chang, J.-F., Kirn, D.H., Liu, T.-C., 2009. Targeted genetic and viral therapy for advanced head and neck cancers. Drug Discov. Today 14, 570–578.

Hubbell, E., Liu, W.-M., Mei, R., 2002. Robust estimators for expression analysis. Bioinformatics 18, 1585–1592.

Hughes, S.A., Carothers, A.M., Hunt, D.H., Moran, A.E., Mueller, J.D., Bertagnolli, M.M., 2002. Adenomatous polyposis coli truncation alters cytoskeletal structure and microtubule stability in early intestinal tumorigenesis. J. Gastrointest. Surg. 6, 868–874; discussion 875.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P., 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31, e15.

Irizarry, R.A., Wang, C., Zhou, Y., Speed, T.P., 2009. Gene set enrichment analysis made simple. Stat Methods Med Res 18, 565–575.

Jankovic, V., Ciarrocchi, A., Boccuni, P., DeBlasio, T., Benezra, R., Nimer, S.D., 2007. Id1 restrains myeloid commitment, maintaining the self-renewal capacity of hematopoietic stem cells. Proc. Natl. Acad. Sci. U.S.A. 104, 1260–1265.

Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D., 2011. Global cancer statistics. CA Cancer J Clin 61, 69–90.

Jensen, D.E., Proctor, M., Marquis, S.T., Gardner, H.P., Ha, S.I., Chodosh, L.A., Ishov, A.M., Tommerup, N., Vissing, H., Sekido, Y., Minna, J., Borodovsky, A., Schultz, D.C., Wilkinson, K.D., Maul, G.G., Barlev, N., Berger, S.L., Prendergast, G.C., Rauscher, F.J., 3rd, 1998. BAP1: a novel ubiquitin hydrolase which binds to the BRCA1 RING finger and enhances BRCA1-mediated cell growth suppression. Oncogene 16, 1097–1112.

Johannessen, C.M., Boehm, J.S., Kim, S.Y., Thomas, S.R., Wardwell, L., Johnson, L.A., Emery, C.M., Stransky, N., Cogdill, A.P., Barretina, J., Caponigro, G., Hieronymus, H., Murray, R.R., Salehi-Ashtiani, K., Hill, D.E., Vidal, M., Zhao, J.J., Yang, X., Alkan, O., Kim, S., Harris, J.L., Wilson, C.J., Myer, V.E., Finan, P.M., Root, D.E., Roberts, T.M., Golub, T., Flaherty, K.T., Dummer, R., Weber, B.L., Sellers, W.R., Schlegel, R., Wargo, J.A., Hahn, W.C., Garraway, L.A., 2010. COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. Nature 468, 968–972.

Jones, P.A., Baylin, S.B., 2002. The fundamental role of epigenetic events in cancer. Nat. Rev. Genet. 3, 415–428.

Jones, R.G., Thompson, C.B., 2009. Tumor suppressors and cell metabolism: a recipe for cancer growth. Genes Dev. 23, 537–548.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M., 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109–114.

Kato, S., Han, S.-Y., Liu, W., Otsuka, K., Shibata, H., Kanamaru, R., Ishioka, C., 2003. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc. Natl. Acad. Sci. U.S.A. 100, 8424–8429.

Kim, S.-Y., Volsky, D.J., 2005. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 6, 144.

Kitagawa, Y., Inoue, K., Sasaki, S., Hayashi, Y., Matsuo, Y., Lieber, M.R., Mizoguchi, H., Yokota, J., Kohno, T., 2002. Prevalent involvement of illegitimate V(D)J recombination in chromosome 9p21 deletions in lymphoid leukemia. J. Biol. Chem. 277, 46289–46297.

Kleinerman, R.A., Tucker, M.A., Tarone, R.E., Abramson, D.H., Seddon, J.M., Stovall, M., Li, F.P., Fraumeni, J.F., Jr, 2005. Risk of new cancers after radiotherapy in long-term survivors of retinoblastoma: an extended follow-up. J. Clin. Oncol. 23, 2272–2279.

Knudson, A.G., Jr, 1971. Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U.S.A. 68, 820–823.

Konopka, J.B., Watanabe, S.M., Witte, O.N., 1984. An alteration of the human c-abl protein in K562 leukemia cells unmasks associated tyrosine kinase activity. Cell 37, 1035–1042.

Kralovics, R., Passamonti, F., Buser, A.S., Teo, S.-S., Tiedt, R., Passweg, J.R., Tichelli, A., Cazzola, M., Skoda, R.C., 2005. A gain-of-function mutation of JAK2 in myeloproliferative disorders. N. Engl. J. Med. 352, 1779–1790.

Kubbutat, M.H., Jones, S.N., Vousden, K.H., 1997. Regulation of p53 stability by Mdm2. Nature 387, 299–303.

Kummar, S., Chen, A., Parchment, R.E., Kinders, R.J., Ji, J., Tomaszewski, J.E., Doroshow, J.H., 2012. Advances in using PARP inhibitors to treat cancer. BMC Medicine 10, 25.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., Lee, D., 2008. Inferring pathway activity toward precise disease classification. PLoS Comput. Biol. 4, e1000217.

Levine, D.M., Haynor, D.R., Castle, J.C., Stepaniants, S.B., Pellegrini, M., Mao, M., Johnson, J.M., 2006. Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. Genome Biol. 7, R93.

Levine, R.L., Wadleigh, M., Cools, J., Ebert, B.L., Wernig, G., Huntly, B.J.P., Boggon, T.J., Wlodarska, I., Clark, J.J., Moore, S., Adelsperger, J., Koo, S., Lee, J.C., Gabriel, S., Mercher, T., D’Andrea, A., Fröhling, S., Döhner, K., Marynen, P., Vandenberghe, P., Mesa, R.A., Tefferi, A., Griffin, J.D., Eck, M.J., Sellers, W.R., Meyerson, M., Golub, T.R., Lee, S.J., Gilliland, D.G., 2005. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Cancer Cell 7, 387–397.

Levy, C., Khaled, M., Fisher, D.E., 2006. MITF: master regulator of melanocyte development and melanoma oncogene. Trends Mol Med 12, 406–414.

Liao, W., Lin, J.-X., Leonard, W.J., 2013. Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. Immunity 38, 13–25.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P., 2011. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740.

Liu, X., Ory, V., Chapman, S., Yuan, H., Albanese, C., Kallakury, B., Timofeeva, O.A., Nealon, C., Dakic, A., Simic, V., Haddad, B.R., Rhim, J.S., Dritschilo, A., Riegel, A., McBride, A., Schlegel, R., 2012. ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. Am. J. Pathol. 180, 599–607.

Liu, X., Shi, Y., Maag, D.X., Palma, J.P., Patterson, M.J., Ellis, P.A., Surber, B.W., Ready, D.B., Soni, N.B., Ladror, U.S., Xu, A.J., Iyer, R., Harlan, J.E., Solomon, L.R., Donawho, C.K., Penning, T.D., Johnson, E.F., Shoemaker, A.R., 2012. Iniparib nonselectively modifies cysteine-containing proteins in tumor cells and is not a bona fide PARP inhibitor. Clin. Cancer Res. 18, 510–523.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E.L., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 14, 1675–1680.

Ma, Y., Dobbins, S.E., Sherborne, A.L., Chubb, D., Galbiati, M., Cazzaniga, G., Micalizzi, C., Tearle, R., Lloyd, A.L., Hain, R., Greaves, M., Houlston, R.S., 2013. Developmental timing of mutations revealed by whole-genome sequencing of twins with acute lymphoblastic leukemia. Proc. Natl. Acad. Sci. U.S.A. 110, 7429–7433.

MacConaill, L.E., Campbell, C.D., Kehoe, S.M., Bass, A.J., Hatton, C., Niu, L., Davis, M., Yao, K., Hanna, M., Mondal, C., Luongo, L., Emery, C.M., Baker, A.C., Philips, J., Goff, D.J., Fiorentino, M., Rubin, M.A., Polyak, K., Chan, J., Wang, Y., Fletcher, J.A., Santagata, S., Corso, G., Roviello, F., Shivdasani, R., Kieran, M.W., Ligon, K.L., Stiles, C.D., Hahn, W.C., Meyerson, M.L., Garraway, L.A., 2009. Profiling critical cancer gene mutations in clinical tumor samples. PLoS ONE 4, e7887.

Martins, C.P., Brown-Swigart, L., Evan, G.I., 2006. Modeling the therapeutic efficacy of p53 restoration in tumors. Cell 127, 1323–1334.

Matsumoto, A., Masuhara, M., Mitsui, K., Yokouchi, M., Ohtsubo, M., Misawa, H., Miyajima, A., Yoshimura, A., 1997. CIS, a cytokine inducible SH2 protein, is a target of the JAK-STAT5 pathway and modulates STAT5 activation. Blood 89, 3148–3154.

Maude, S.L., Tasian, S.K., Vincent, T., Hall, J.W., Sheen, C., Roberts, K.G., Seif, A.E., Barrett, D.M., Chen, I.-M., Collins, J.R., Mullighan, C.G., Hunger, S.P., Harvey, R.C., Willman, C.L., Fridman, J.S., Loh, M.L., Grupp, S.A., Teachey, D.T., 2012. Targeting JAK1/2 and mTOR in murine xenograft models of Ph-like acute lymphoblastic leukemia. Blood 120, 3510–3518.

McKenna, E.S., Sansam, C.G., Cho, Y.-J., Greulich, H., Evans, J.A., Thom, C.S., Moreau, L.A., Biegel, J.A., Pomeroy, S.L., Roberts, C.W.M., 2008. Loss of the epigenetic tumor suppressor SNF5 leads to cancer without genomic instability. Mol. Cell. Biol. 28, 6223–6233.

McPherson, A., Wu, C., Hajirasouliha, I., Hormozdiari, F., Hach, F., Lapuk, A., Volik, S., Shah, S., Collins, C., Sahinalp, S.C., 2011. Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. Bioinformatics 27, 1481–1488.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., Getz, G., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, R41.

Mitchell, T.J., Whittaker, S.J., John, S., 2003. Dysregulated expression of COOH-terminally truncated Stat5 and loss of IL2-inducible Stat5-dependent gene expression in Sezary Syndrome. Cancer Res. 63, 9048–9054.

Mizuki, M., Fenski, R., Halfter, H., Matsumura, I., Schmidt, R., Müller, C., Grüning, W., Kratz-Albers, K., Serve, S., Steur, C., Büchner, T., Kienast, J., Kanakura, Y., Berdel, W.E., Serve, H., 2000. Flt3 mutations from patients with acute myeloid leukemia induce transformation of 32D cells mediated by the Ras and STAT5 pathways. Blood 96, 3907–3914.

Monks, A., Scudiero, D., Skehan, P., Shoemaker, R., Paull, K., Vistica, D., Hose, C., Langley, J., Cronise, P., Vaigro-Wolff, A., 1991. Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. J. Natl. Cancer Inst. 83, 757–766.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., Groop, L.C., 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34, 267–273.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628.

Muchardt, C., Yaniv, M., 1999. The mammalian SWI/SNF complex and the control of cell growth. Semin. Cell Dev. Biol. 10, 189–195.

Murray, P.J., 2007. The JAK-STAT signaling pathway: input and output integration. J. Immunol. 178, 2623–2629.

Nazarian, R., Shi, H., Wang, Q., Kong, X., Koya, R.C., Lee, H., Chen, Z., Lee, M.-K., Attar, N., Sazegar, H., Chodon, T., Nelson, S.F., McArthur, G., Sosman, J.A., Ribas, A., Lo, R.S., 2010. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. Nature 468, 973–977.

Nevins, J.R., 2001. The Rb/E2F pathway and cancer. Hum. Mol. Genet. 10, 699–703.

Nieborowska-Skorska, M., Slupianek, A., Xue, L., Zhang, Q., Raghunath, P.N., Hoser, G., Wasik, M.A., Morris, S.W., Skorski, T., 2001. Role of signal transducer and activator of transcription 5 in nucleophosmin/ anaplastic lymphoma kinase-mediated malignant transformation of lymphoid cells. Cancer Res. 61, 6517–6523.

Norton, J.D., Deed, R.W., Craggs, G., Sablitzky, F., 1998. Id helix-loop-helix proteins in cell growth and differentiation. Trends Cell Biol. 8, 58–65.

Nowell, P.C., Hungerford, D.A., 1961. Chromosome studies in human leukemia. II. Chronic granulocytic leukemia. J. Natl. Cancer Inst. 27, 1013–1035.

Nurtdinov, R.N., Vasiliev, M.O., Ershova, A.S., Lossev, I.S., Karyagina, A.S., 2010. PLANdbAffy: probe-level annotation database for Affymetrix expression microarrays. Nucleic Acids Res. 38, D726–730.

Nussbaum, R.L., McInnes, Willard, H.F., Hamosh, A., Thompson, M.W., 2007. Thompson & Thompson genetics in medicine [WWW Document]. URL http://www.mdconsult.com/public/book/view?title=Nussbaum:+Thompson+&+Thompson+Genetics+in+Medicine

Nyga, A., Cheema, U., Loizidou, M., 2011. 3D tumour models: novel in vitro approaches to cancer studies. J Cell Commun Signal 5, 239–248.

Olivier, M., Goldgar, D.E., Sodha, N., Ohgaki, H., Kleihues, P., Hainaut, P., Eeles, R.A., 2003. Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype. Cancer Res. 63, 6643–6650.

Ooi, H.S., Schneider, G., Lim, T.-T., Chan, Y.-L., Eisenhaber, B., Eisenhaber, F., 2010. Biomolecular pathway databases. Methods Mol. Biol. 609, 129–144.

Pandita, A., Aldape, K.D., Zadeh, G., Guha, A., James, C.D., 2004. Contrasting in vivo and in vitro fates of glioblastoma cell subpopulations with amplified EGFR. Genes Chromosomes Cancer 39, 29–36.

Patel, A.G., De Lorenzo, S.B., Flatten, K.S., Poirier, G.G., Kaufmann, S.H., 2012. Failure of iniparib to inhibit poly(ADP-Ribose) polymerase in vitro. Clin. Cancer Res. 18, 1655–1662.

Pavlidis, P., Lewis, D.P., Noble, W.S., 2002. Exploring gene expression data with class scores. Pac Symp Biocomput 474–485.

Payne, S.R., Kemp, C.J., 2005. Tumor suppressor genetics. Carcinogenesis 26, 2031–2045.

Peña-Llopis, S., Vega-Rubín-de-Celis, S., Liao, A., Leng, N., Pavía-Jiménez, A., Wang, S., Yamasaki, T., Zhrebker, L., Sivanand, S., Spence, P., Kinch, L., Hambuch, T., Jain, S., Lotan, Y., Margulis, V., Sagalowsky, A.I., Summerour, P.B., Kabbani, W., Wong, S.W.W., Grishin, N., Laurent, M., Xie, X.-J., Haudenschild, C.D., Ross, M.T., Bentley, D.R., Kapur, P., Brugarolas, J., 2012. BAP1 loss defines a new class of renal cell carcinoma. Nat. Genet. 44, 751–759.

Pepper, S.D., Saunders, E.K., Edwards, L.E., Wilson, C.L., Miller, C.J., 2007. The utility of MAS5 expression summary and detection call algorithms. BMC Bioinformatics 8, 273.

Perry, S.S., Zhao, Y., Nie, L., Cochrane, S.W., Huang, Z., Sun, X.-H., 2007. Id1, but not Id3, directs long-term repopulating hematopoietic stem-cell maintenance. Blood 110, 2351–2360.

Pesu, M., Laurence, A., Kishore, N., Zwillich, S.H., Chan, G., O’Shea, J.J., 2008. Therapeutic targeting of Janus kinases. Immunol. Rev. 223, 132–142.

Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S.V., Hainaut, P., Olivier, M., 2007. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum. Mutat. 28, 622–629.

Pikman, Y., Lee, B.H., Mercher, T., McDowell, E., Ebert, B.L., Gozo, M., Cuker, A., Wernig, G., Moore, S., Galinsky, I., DeAngelo, D.J., Clark, J.J., Lee, S.J., Golub, T.R., Wadleigh, M., Gilliland, D.G., Levine, R.L., 2006. MPLW515L is a novel somatic activating mutation in myelofibrosis with myeloid metaplasia. PLoS Med. 3, e270.

Ponder, B.A., 2001. Cancer genetics. Nature 411, 336–341.

Quintás-Cardama, A., Cortes, J.E., 2006. Chronic myeloid leukemia: diagnosis and treatment. Mayo Clin. Proc. 81, 973–988.

Ram, P.A., Waxman, D.J., 1999. SOCS/CIS protein inhibition of growth hormone-stimulated STAT5 signaling by multiple mechanisms. J. Biol. Chem. 274, 35553–35561.

Rawlings, J.S., Rosler, K.M., Harrison, D.A., 2004. The JAK/STAT signaling pathway. J. Cell. Sci. 117, 1281–1283.

Ray-Coquard, I., Blay, J.-Y., Italiano, A., Le Cesne, A., Penel, N., Zhi, J., Heil, F., Rueger, R., Graves, B., Ding, M., Geho, D., Middleton, S.A., Vassilev, L.T., Nichols, G.L., Bui, B.N., 2012. Effect of the MDM2 antagonist RG7112 on the P53 pathway in patients with MDM2-amplified, well-differentiated or dedifferentiated liposarcoma: an exploratory proof-of-mechanism study. Lancet Oncol. 13, 1133–1140.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26.

Robles, A.I., Linke, S.P., Harris, C.C., 2002. The p53 network in lung carcinogenesis. Oncogene 21, 6898–6907.

Rowley, J.D., 1973. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 243, 290–293.

Ryan, K.M., Phillips, A.C., Vousden, K.H., 2001. Regulation and function of the p53 tumor suppressor protein. Curr. Opin. Cell Biol. 13, 332–337.

Sasaki, S., Kitagawa, Y., Sekido, Y., Minna, J.D., Kuwano, H., Yokota, J., Kohno, T., 2003. Molecular processes of chromosome 9p21 deletions in human cancers. Oncogene 22, 3792–3798.

Scheithauer, W., Temsch, E.M., Moyer, M.P., Grabner, G., 1987. Search for improved culture conditions for clonogenic growth of human colorectal cancer cells in vitro. Int. J. Cell Cloning 5, 55–70.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

Scherer, W.F., Syverton, J.T., Gey, G.O., 1953. Studies on the propagation in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. J. Exp. Med. 97, 695–710.

Scott, L.M., Tong, W., Levine, R.L., Scott, M.A., Beer, P.A., Stratton, M.R., Futreal, P.A., Erber, W.N., McMullin, M.F., Harrison, C.N., Warren, A.J., Gilliland, D.G., Lodish, H.F., Green, A.R., 2007. JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. N. Engl. J. Med. 356, 459–468.

Ségalat, L., 2007. Loss-of-function genetic diseases and the concept of pharmaceutical targets. Orphanet J Rare Dis 2, 30.

Sen, B., Peng, S., Woods, D.M., Wistuba, I., Bell, D., El-Naggar, A.K., Lai, S.Y., Johnson, F.M., 2012. STAT5A-mediated SOCS2 expression regulates Jak2 and STAT3 activity following c-Src inhibition in head and neck squamous carcinoma. Clin. Cancer Res. 18, 127–139.

Shah, N.P., Nicoll, J.M., Nagar, B., Gorre, M.E., Paquette, R.L., Kuriyan, J., Sawyers, C.L., 2002. Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. Cancer Cell 2, 117–125.

Shah, N.P., Tran, C., Lee, F.Y., Chen, P., Norris, D., Sawyers, C.L., 2004. Overriding imatinib resistance with a novel ABL kinase inhibitor. Science 305, 399–401.

Shain, A.H., Pollack, J.R., 2013. The spectrum of SWI/SNF mutations, ubiquitous in human cancers. PLoS ONE 8, e55119.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311.

Shoemaker, R.H., Monks, A., Alley, M.C., Scudiero, D.A., Fine, D.L., McLemore, T.L., Abbott, B.J., Paull, K.D., Mayo, J.G., Boyd, M.R., 1988. Development of human tumor cell line panels for use in disease-oriented drug screening. Prog. Clin. Biol. Res. 276, 265–286.

Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3, Article3.

Sobel, R.E., Sadar, M.D., 2005. Cell lines used in prostate cancer research: a compendium of old and new lines--part 1. J. Urol. 173, 342–359.

Solimini, N.L., Xu, Q., Mermel, C.H., Liang, A.C., Schlabach, M.R., Luo, J., Burrows, A.E., Anselmo, A.N., Bredemeyer, A.L., Li, M.Z., Beroukhim, R., Meyerson, M., Elledge, S.J., 2012. Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science 337, 104–109.

Sonkin, D., Hassan, M., Murphy, D.J., Tatarinova, T.V., 2013. Tumor suppressors status in cancer cell line encyclopedia. Mol Oncol 7, 791–798.

Stanbridge, E.J., 1990. Human Tumor Suppressor Genes. Annual Review of Genetics 24, 615–657.

Stark, G.R., Darnell, J.E., Jr, 2012. The JAK-STAT pathway at twenty. Immunity 36, 503–514.

Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U.S.A. 100, 9440–9445.

Strachan, T., Read, A.P., Strachan, 2011. Human molecular genetics. Garland Science, New York.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 102, 15545–15550.

Sun, H., Chang, Y., Schweers, B., Dyer, M.A., Zhang, X., Hayward, S.W., Goodrich, D.W., 2006. An E2F binding-deficient Rb1 protein partially rescues developmental defects associated with Rb1 nullizygosity. Mol. Cell. Biol. 26, 1527–1537.

Tamayo, P., Steinhardt, G., Liberzon, A., Mesirov, J.P., 2012. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res.

Tap, W.D., Gong, K.-W., Dering, J., Tseng, Y., Ginther, C., Pauletti, G., Glaspy, J.A., Essner, R., Bollag, G., Hirth, P., Zhang, C., Slamon, D.J., 2010. Pharmacodynamic characterization of the efficacy signals due to selective BRAF inhibition with PLX4032 in malignant melanoma. Neoplasia 12, 637–649.

Tentler, J.J., Tan, A.C., Weekes, C.D., Jimeno, A., Leong, S., Pitts, T.M., Arcaroli, J.J., Messersmith, W.A., Eckhardt, S.G., 2012. Patient-derived tumour xenografts as models for oncology drug development. Nat Rev Clin Oncol 9, 338–350.

Testa, J.R., Cheung, M., Pei, J., Below, J.E., Tan, Y., Sementino, E., Cox, N.J., Dogan, A.U., Pass, H.I., Trusa, S., Hesdorffer, M., Nasu, M., Powers, A., Rivera, Z., Comertpay, S., Tanji, M., Gaudino, G., Yang, H., Carbone, M., 2011. Germline BAP1 mutations predispose to malignant mesothelioma. Nat. Genet. 43, 1022–1025.

Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S., Park, P.J., 2005. Discovering statistically significant pathways in expression profiling studies. Proc. Natl. Acad. Sci. U.S.A. 102, 13544–13549.

Tomfohr, J., Lu, J., Kepler, T.B., 2005. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics 6, 225.

Tsai, J., Lee, J.T., Wang, W., Zhang, J., Cho, H., Mamo, S., Bremer, R., Gillette, S., Kong, J., Haass, N.K., Sproesser, K., Li, L., Smalley, K.S.M., Fong, D., Zhu, Y.-L., Marimuthu, A., Nguyen, H., Lam, B., Liu, J., Cheung, I., Rice, J., Suzuki, Y., Luu, C., Settachatgul, C., Shellooe, R., Cantwell, J., Kim, S.-H., Schlessinger, J., Zhang, K.Y.J., West, B.L., Powell, B., Habets, G., Zhang, C., Ibrahim, P.N., Hirth, P., Artis, D.R., Herlyn, M., Bollag, G., 2008. Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. Proc. Natl. Acad. Sci. U.S.A. 105, 3041–3046.

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U.S.A. 98, 5116–5121.

Ungefroren, H., Sebens, S., Seidl, D., Lehnert, H., Hass, R., 2011. Interaction of tumor cells with the microenvironment. Cell Commun. Signal 9, 18.

Vainchenker, W., Delhommeau, F., Constantinescu, S.N., Bernard, O.A., 2011. New mutations and pathogenesis of myeloproliferative neoplasms. Blood 118, 1723–1735.

Vassilev, L.T., 2004. In Vivo Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2. Science 303, 844–848.

Ventii, K.H., Devi, N.S., Friedrich, K.L., Chernova, T.A., Tighiouart, M., Van Meir, E.G., Wilkinson, K.D., 2008. BRCA1-associated protein-1 is a tumor suppressor that requires deubiquitinating activity and nuclear localization. Cancer Res. 68, 6953–6962.

Ventura, A., Kirsch, D.G., McLaughlin, M.E., Tuveson, D.A., Grimm, J., Lintault, L., Newman, J., Reczek, E.E., Weissleder, R., Jacks, T., 2007. Restoration of p53 function leads to tumour regression in vivo. Nature 445, 661–665.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Kinzler, K.W., 2013. Cancer Genome Landscapes. Science 339, 1546–1558.

Vousden, K.H., Lu, X., 2002. Live or let die: the cell’s response to p53. Nat. Rev. Cancer 2, 594–604.

Wagle, N., Emery, C., Berger, M.F., Davis, M.J., Sawyer, A., Pochanard, P., Kehoe, S.M., Johannessen, C.M., Macconaill, L.E., Hahn, W.C., Meyerson, M., Garraway, L.A., 2011. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. J. Clin. Oncol. 29, 3085–3096.

Wang, Z., Bhattacharya, N., Weaver, M., Petersen, K., Meyer, M., Gapter, L., Magnuson, N.S., 2001. Pim-1: a serine/threonine kinase with a role in cell survival, proliferation, differentiation and tumorigenesis. J. Vet. Sci. 2, 167–179.

van Waveren, C., Moraes, C.T., 2008. Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system. BMC Genomics 9, 18.

Weinberg, R.A., 2007. The Biology of cancer. Garland Science, New York.

Weinstein, J.N., Myers, T.G., O’Connor, P.M., Friend, S.H., Fornace, A.J., Jr, Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., Buolamwini, J.K., van Osdol, W.W., Monks, A.P., Scudiero, D.A., Sausville, E.A., Zaharevitz, D.W., Bunow, B., Viswanadhan, V.N., Johnson, G.S., Wittes, R.E., Paull, K.D., 1997. An information-intensive approach to the molecular pharmacology of cancer. Science 275, 343–349.

Weisberg, E., Manley, P.W., Breitenstein, W., Brüggen, J., Cowan-Jacob, S.W., Ray, A., Huntly, B., Fabbro, D., Fendrich, G., Hall-Meyers, E., Kung, A.L., Mestan, J., Daley, G.Q., Callahan, L., Catley, L., Cavazza, C., Azam, M., Mohammed, A., Neuberg, D., Wright, R.D., Gilliland, D.G., Griffin, J.D., 2005. Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. Cancer Cell 7, 129–141.

White, E., 2003. The pims and outs of survival signaling: role for the Pim-2 protein kinase in the suppression of apoptosis by cytokines. Genes Dev. 17, 1813–1816.

Wick, M.R., Ritter, J.H., Dehner, L.P., 1995. Malignant rhabdoid tumors: a clinicopathologic review and conceptual discussion. Semin Diagn Pathol 12, 233–248.

Wood, A.D., Chen, E., Donaldson, I.J., Hattangadi, S., Burke, K.A., Dawson, M.A., Miranda-Saavedra, D., Lodish, H.F., Green, A.R., Göttgens, B., 2009. ID1 promotes expansion and survival of primary erythroid cells and is a target of JAK2V617F-STAT5 signaling. Blood 114, 1820–1830.

Wu, J.N., Roberts, C.W.M., 2013. ARID1A mutations in cancer: another epigenetic tumor suppressor? Cancer Discov 3, 35–43.

Wu, X., Hepner, K., Castelino-Prabhu, S., Do, D., Kaye, M.B., Yuan, X.J., Wood, J., Ross, C., Sawyers, C.L., Whang, Y.E., 2000. Evidence for regulation of the PTEN tumor suppressor by a membrane-localized multi-PDZ domain containing scaffold protein MAGI-2. Proc. Natl. Acad. Sci. U.S.A. 97, 4233–4238.

Xue, W., Kitzing, T., Roessler, S., Zuber, J., Krasnitz, A., Schultz, N., Revill, K., Weissmueller, S., Rappaport, A.R., Simon, J., Zhang, J., Luo, W., Hicks, J., Zender, L., Wang, X.W., Powers, S., Wigler, M., Lowe, S.W., 2012. A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. Proc. Natl. Acad. Sci. U.S.A. 109, 8212–8217.

Xue, W., Zender, L., Miething, C., Dickins, R.A., Hernando, E., Krizhanovsky, V., Cordon-Cardo, C., Lowe, S.W., 2007. Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. Nature 445, 656–660.

Yang, Z., Wang, D., Wang, G., Zhang, Q., Liu, J., Peng, P., Liu, X., 2010. Clinical study of recombinant adenovirus-p53 combined with fractionated stereotactic radiotherapy for hepatocellular carcinoma. J. Cancer Res. Clin. Oncol. 136, 625–630.

Zajac, V., Tomka, M., Ilenciková, D., Májek, P., Stevurková, V., Kirchhoff, T., 2000. A double germline mutations in the APC and p53 genes. Neoplasma 47, 335–341.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C., Weinstein, J.N., 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 4, R28.

Zhang, S., Fukuda, S., Lee, Y., Hangoc, G., Cooper, S., Spolski, R., Leonard, W.J., Broxmeyer, H.E., 2000. Essential role of signal transducer and activator of transcription (Stat)5a but not Stat5b for Flt3-dependent signaling. J. Exp. Med. 192, 719–728.

Zhong, S., Storch, K.-F., Lipan, O., Kao, M.-C.J., Weitz, C.J., Wong, W.H., 2004. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. Appl. Bioinformatics 3, 261–264.

Zhou, X., Su, Z., 2007. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics 8, 246.

Appendix A1.1 - D. Sonkin publications

First author publications:

Sonkin, D., Hassan, M., Murphy, D.J., Tatarinova, T.V., 2013. Tumor Suppressors Status in Cancer Cell Line Encyclopedia. Molecular Oncology Volume 7, Issue 4, Pages 791–798

Sonkin, D., Palmer, M., Rong, X., Horrigan, K., Regnier, C., Fanton, C., Holash, J., Pinzon-Ortiz, M., Squires, M., Sirulnik, A., Radimerski, T., Schlegel, R., Morrissey, M., Cao, A., pSTAT5 mRNA expression signature in hematopoietic cancer cell lines. Manuscript under review.

Contributing author publications:

Abazeed, M.E., Adams, D.J., Hurov, K.E., Tamayo, P., Creighton, C.J., **Sonkin, D.,** Giacomelli, A.O., Du, C., Fries, D.F., Wong, K.-K., Mesirov, J.P., Loeffler, J.S., Schreiber, S.L., Hammerman, P.S., Meyerson, M., 2013. Integrative radiogenomic profiling of squamous cell lung cancer. AACR Cancer Research. doi: 10.1158/0008-5472.CAN-13-1616

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G., **Sonkin, D.,** Reddy, A., et al., 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607.

Scholer-Dahirel, A., Schlabach, M.R., Loo, A., Bagdasarian, L., Meyer, R., Guo, R., Woolfenden, S., Yu, K.K., Markovits, J., Killary, K., **Sonkin, D.,** Yao, Y.-M., Warmuth, M., Sellers, W.R., Schlegel, R., Stegmeier, F., Mosher, R.E., McLaughlin, M.E., 2011. Maintenance of adenomatous polyposis coli (APC)-mutant colorectal cancer is dependent on Wnt/beta-catenin signaling. Proc. Natl. Acad. Sci. U.S.A. 108, 17135–17140.

Venkatesan, K., Stransky, N., Margolin, A.A., Reddy, A., Raman, P., **Sonkin, D.,** Jones, M.D., Wilson, C.J., Kim, S., Warmuth, M., Sellers, W.R., Lehár, J., Barretina, J., Caponigro, G., Garraway, L.A., Morrissey, M.P., 2010. Computational prediction of compound sensitivity with genomic signatures. Proceedings of the American Association for Cancer Research.

Kuraguchi, M., Ohene-Baah, N.Y., **Sonkin, D.,** Bronson, R.T., Kucherlapati, R., 2009. Genetic mechanisms in Apc-mediated mammary tumorigenesis. PLoS Genet. 5, e1000367.

Alterovitz, G., Benson, R., Ramoni, M.F., 2009. Automation in proteomics and genomics an engineering case-based approach. John Wiley, Chichester, West Sussex, U.K.; Hoboken, N.J., **book chapter co-author**.

Kucherlapati, M.H., Yang, K., Fan, K., Kuraguchi, M., **Sonkin, D.,** Rosulek, A., Lipkin, M., Bronson, R.T., Aronow, B.J., Kucherlapati, R., 2008. Loss of Rb1 in the gastrointestinal tract of Apc1638N mice promotes tumors of the cecum and proximal colon. Proc. Natl. Acad. Sci. U.S.A. 105, 15493–15498.

Appendix A1.2 List of 69 well-known and putative tumor suppressor genes.

Gene	ENTREZ ID	Gene Title
APC	324	adenomatous polyposis coli
ARID1A	8289	AT rich interactive domain 1A (SWI-like)
ATM	472	ataxia telangiectasia mutated
ATR	545	ataxia telangiectasia and Rad3 related
BMPR1A	657	bone morphogenetic protein receptor, type IA
BRCA1	672	breast cancer 1, early onset
BRCA2	675	breast cancer 2, early onset
BRIP1	83990	BRCA1 interacting protein C-terminal helicase 1
CDC73	79577	cell division cycle 73, Paf1/RNA polymerase II complex component
CDH1	999	cadherin 1, type 1, E-cadherin (epithelial)
CDKN1A	1026	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
CDKN1B	1027	cyclin-dependent kinase inhibitor 1B (p27, Kip1)
CDKN2A	1029	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
CDKN2B	1030	cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)
CHEK1	1111	Checkpoint kinase Chk1 (CHK1)
CHEK2	11200	CHK2 checkpoint homolog (S. pombe)
CREBBP	1387	CREB binding protein
CYLD	1540	cylindromatosis (turban tumor syndrome)
DLC1	10395	deleted in liver cancer 1
FANCA	2175	Fanconi anemia, complementation group A
FANCB	2187	Fanconi anemia, complementation group B
FANCC	2176	Fanconi anemia, complementation group C
FANCD2	2177	Fanconi anemia, complementation group D2
FANCE	2178	Fanconi anemia, complementation group E
FANCF	2188	Fanconi anemia, complementation group F
FANCG	2189	Fanconi anemia, complementation group G
FANCI	55215	Fanconi anemia, complementation group I
FANCL	55120	Fanconi anemia, complementation group L
FANCM	57697	Fanconi anemia, complementation group M
FBXW7	55294	F-box and WD repeat domain containing 7
FH	2271	fumarate hydratase
FHIT	2272	fragile histidine triad gene
FLCN	201163	folliculin
HIPK2	28996	homeodomain interacting protein kinase 2
KDM6A	7403	ubiquitously transcribed tetratricopeptide repeat, X chromosome
LATS1	9113	LATS, large tumor suppressor, homolog 1 (Drosophila)
LATS2	26524	LATS, large tumor suppressor, homolog 2 (Drosophila)
MEN1	4221	multiple endocrine neoplasia I

MLH1	4292	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
MSH2	4436	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)
MSH6	2956	mutS homolog 6
NBN	4683	nibrin
NF1	4763	neurofibromin 1
NF2	4771	neurofibromin 2 (merlin)
PALB2	79728	partner and localizer of BRCA2
PRKAR1A	5573	protein kinase, cAMP-dependent, regulatory, type I, alpha
PTCH1	5727	patched homolog 1 (Drosophila)
PTEN	5728	Putative protein tyrosine phosphatase (PTEN)
RAD50	10111	RAD50 homolog (S. cerevisiae)
RAD51	5888	RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae)
RB1	5925	retinoblastoma 1
RUNX1	861	runt-related transcription factor 1
SDHB	6390	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
SDHD	6392	succinate dehydrogenase complex, subunit D
SMAD2	4087	SMAD family member 2
SMAD4	4089	SMAD family member 4
SMARCB1	6598	SWI/SNF related, actin dependent regulator of chromatin, subfamily b 1
STK11	6794	serine/threonine kinase 11
STK3	6788	serine/threonine kinase 3 (STE20 homolog, yeast)
SUFU	51684	suppressor of fused homolog (Drosophila)
TGFBR2	7048	transforming growth factor, beta receptor II (70/80kDa)
TNFAIP3	7128	tumor necrosis factor, alpha-induced protein 3
TP53	7157	tumor protein p53
TP53BP1	7158	tumor protein p53 binding protein
TSC1	7248	tuberous sclerosis 1
TSC2	7249	tuberous sclerosis 2
VHL	7428	von Hippel-Lindau tumor suppressor
WRN	7486	Werner syndrome
WT1	7490	Wilms tumor 1

Appendix A1.3 Summary of inactivation categories counts for 69 tumor suppressor genes

Gene	G	E-G		G-M	G-D		E-G-D	E-G-M		E	E-LOH		WT-E	WT
Symbol	#	#		#	#		#	#		#	#		#	#
CDKN2A	231	16		0	231		16	0		42	15		97	101
CDKN2B	215	0		0	215		0	0		0	0		18	218
TP53	179	39		176	3		21	18		0	37		84	89
RB1	36	5		29	7		3	2		0	7		295	5
PTEN	20	9		14	6		9	0		2	7		398	0
SMAD4	20	0		6	14		0	0		0	0		6	353
KDM6A	15	0		1	14		0	0		0	0		0	132
APC	13	2		13	0		2	0		0	1		226	141
NF1	10	0		7	3		0	0		0	0		13	419
TGFBR2	8	21		0	8		15	6		52	30		235	96
MLH1	7	1		6	1		0	1		18	4		350	2
FHIT	6	0		0	6		0	0		0	0		5	232
CDH1	5	0		2	3		0	0		0	0		183	234
STK11	5	0		3	2		0	0		0	0		0	414
NF2	4	6		3	1		6	0		0	3		224	197
MSH2	4	5		2	2		4	1		0	0		519	66
ARID1A	4	2		3	1		1	1		0	2		98	251
DLC1	4	0		2	2		0	0		0	0		103	241
VHL	3	4		3	0		4	0		0	0		13	413
SMARCB1	3	1		0	3		0	1		0	0		288	157
ATM	3	0		3	0		0	0		0	1		169	120
RUNX1	3	0		0	3		0	0		0	0		6	496
LATS2	2	17		0	2		17	0		52	25		214	88
TSC2	2	1		1	1		1	0		1	2		24	427
CREBBP	2	1		1	1		1	0		0	0		426	5
BMPR1A	2	0		0	2		0	0		22	12		301	91
FBXW7	1	2		1	0		2	0		1	1		53	257
CDKN1B	1	1		0	1		1	0		0	0		524	3
BRCA1	1	0		1	0		0	0		1	0		321	78
FANCA	1	0		0	1		0	0		0	0		3	535
FANCB	1	0		0	1		0	0		0	0		0	120
FANCC	1	0		1	0		0	0		0	0		0	418
FANCG	1	0		1	0		0	0		0	0		309	41
FANCM	1	0		1	0		0	0		0	0		0	396
LATS1	1	0		1	0		0	0		0	0		0	422
MSH6	1	0		0	1		0	0		0	0		555	1
PRKAR1A	1	0		0	1		0	0		0	0		1	549

TNFAIP3	1	0		1	0		0	0		0	0		84	327
TSC1	1	0		1	0		0	0		0	0		74	368
WRN	1	0		0	1		0	0		0	0		71	280
PTCH1	0	3		0	0		3	0		48	17		142	212
HIPK2	0	1		0	0		1	0		34	9		228	288
FANCF	0	1		0	0		1	0		10	5		95	375
CDKN1A	0	0		0	0		0	0		41	10		239	275
STK3	0	0		0	0		0	0		6	3		370	120
CHEK2	0	0		0	0		0	0		3	0		74	311
CHEK1	0	0		0	0		0	0		1	0		429	46
FANCD2	0	0		0	0		0	0		1	0		350	43
RAD51	0	0		0	0		0	0		1	0		319	155
ATR	0	0		0	0		0	0		0	0		534	7
BRCA2	0	0		0	0		0	0		0	0		7	295
BRIP1	0	0		0	0		0	0		0	0		13	465
CDC73	0	0		0	0		0	0		0	0		445	53
CYLD	0	0		0	0		0	0		0	0		28	448
FANCE	0	0		0	0		0	0		0	0		2	493
FANCI	0	0		0	0		0	0		0	0		506	10
FANCL	0	0		0	0		0	0		0	0		572	19
FH	0	0		0	0		0	0		0	0		440	133
FLCN	0	0		0	0		0	0		0	0		2	288
MEN1	0	0		0	0		0	0		0	0		261	300
NBN	0	0		0	0		0	0		0	0		211	318
PALB2	0	0		0	0		0	0		0	0		317	186
RAD50	0	0		0	0		0	0		0	0		1	404
SDHB	0	0		0	0		0	0		0	0		458	0
SDHD	0	0		0	0		0	0		0	0		463	0
SMAD2	0	0		0	0		0	0		0	0		375	3
SUFU	0	0		0	0		0	0		0	0		0	420
TP53BP1	0	0		0	0		0	0		0	0		146	155
WT1	0	0		0	0		0	0		0	0		55	467

Appendix A1.4 Lowest GeneGo Gene Set Activity Scores for liver sample from (Ge et al., 2005).

Gene Set Name	Gene Set Size	Gene Set Activity Score	Permutation Fraction
HNRPD_inhibited	1	-1.88	0.998
Development_PDGF signaling via MAPK cascades	46	-1.90	1
Cell cycle_Start of DNA replication in early S phase	43	-1.91	1
Cytoskeleton remodeling_Role PKA in cytoskeleton reorganization	81	-1.93	1
Cytoskeleton remodeling_RalA regulation pathway	33	-1.97	0.998
Transport_ACM3 in salivary glands	55	-1.99	1
wtCFTR and deltaF508 traffic / Membrane expression (norm and CF)	43	-2.00	1
Development_WNT signaling pathway. Part 1. Degradation of beta-catenin in the absence WNT signaling	28	-2.00	1
Translation _Regulation activity of EIF2	56	-2.01	1
Immune response _CCR3 signaling in eosinophils	119	-2.03	1
dGTP metabolism	38	-2.03	1
Gqa specific GPCRs (in brain)	31	-2.07	1
Development_Endothelin-1/EDNRA signaling	69	-2.10	1
Neurophysiological process_PGE2-induced pain processing	39	-2.16	1
Cytoskeleton remodeling_Regulation of actin cytoskeleton by Rho GTPases	62	-2.16	1
Cytoskeleton remodeling_Keratin filaments	48	-2.20	1
Immune response _Function MEF2 in T lymphocytes	92	-2.21	1
Cell adhesion_Integrin-mediated cell adhesion	90	-2.27	1
Cell cycle_Role of Nek in cell cycle regulation	55	-2.29	1
KLF10_inhibited	1	-2.34	1
SKIL_inhibited	1	-2.34	1
Cytoskeleton remodeling_Reverse signalling by ephrin B	80	-2.46	1
Cytoskeleton remodeling_Slit-Robo signaling	56	-2.49	1
Transcription_Role of heterochromatin protein 1 (HP1) family in transcriptional silencing	45	-2.49	1
Transcription_Ligand-Dependent	110	-2.49	1

Transcription of Retinoid-Target genes			
Regulation of CFTR activity (norm and CF)	93	-2.51	1
Apoptosis and survival_BAD phosphorylation	83	-2.55	1
Cell cycle_Spindle assembly and chromosome separation	88	-3.08	1
Cytoskeleton remodeling_Neurofilaments	50	-3.24	1
Development_Role of CDK5 in neuronal development	76	-3.29	1

Appendix A1.5 Fold change and t-test p-values between p-STAT5(+) and p-STAT5(-) cell lines.

Gene Name	Entrez	probe set	p-STAT5+ mean	p-STAT5- mean	Fold	t-test p-value
PIM1	5292	209193_at	875	134	5.04	6.82E-07
CISH	1154	223961_s_at	245	21	4.15	5.86E-06
SOCS2	8835	203373_at	2441	326	6.63	1.64E-05
ID1	3397	208937_s_at	1548	332	4.19	0.00331972
LCN2	3934	212531_at	80	8	2.24	0.00453474
EPOR	2057	209962_at	118	38	1.91	0.00836353
KIR3DL1	3811	211687_x_at	24	14	1.15	0.02315812
C3AR1	719	209906_at	91	35	1.66	0.02897651
BCL2L1	598	212312_at	270	167	1.47	0.03413896
IGJ	3512	212592_at	106	3746	-	0.04997906
EGR1	1958	227404_s_at	1035	351	2.71	0.0638939
OSM	5008	230170_at	53	17	1.55	0.10218279
TBX21	30009	220684_at	40	12	1.46	0.14215803
TNFRSF13B	23495	207641_at	27	71	-1.57	0.15316237
ESR1	2099	205225_at	10	18	-1.15	0.15905403
XIAP	331	228363_at	711	1041	-1.43	0.20670021
ABCB1	5243	243951_at	34	19	1.21	0.21057215
IL18	3606	206295_at	91	50	1.41	0.26985569
SKP2	6502	210567_s_at	256	345	-1.29	0.27693167
MYC	4609	202431_s_at	5556	4662	1.19	0.30379619
SRP9	6726	201273_s_at	5997	6579	-1.1	0.36038668
FOS	2353	209189_at	98	55	1.41	0.42764108
IL10	3586	207433_at	7	23	-1.29	0.45530643
EBF1	1879	227646_at	565	1033	-1.76	0.46111412
CSN1S1	1446	208350_at	4	3	1.02	0.50498373
ONECUT1	3175	210745_at	8	10	-1.03	0.54105495
HSD3B2	3284	206294_at	4	5	-1.02	0.54197609
SLC30A2	7780	230084_at	16	15	1.02	0.54712739
SP1	6667	224760_at	367	311	1.15	0.55826135
PRF1	5551	214617_at	76	73	1.02	0.56205153
IFNG	3458	210354_at	8	9	-1.03	0.56455525
IL22	50616	222974_at	6	4	1.02	0.56529166
CITED4	163732	228625_at	38	94	-1.64	0.58702634
CCND1	595	208712_at	40	107	-1.75	0.60420565
RAD51	5888	205024_s_at	576	626	-1.08	0.66382789
PAX5	5079	206802_at	9	11	-1.03	0.68588032
CSN2	1447	207951_at	10	11	-1.02	0.69349354
SOCS1	8651	210001_s_at	142	102	1.26	0.72728647
RBMS1	5937	225265_at	310	296	1.04	0.7600465
PTGS2	5743	204748_at	28	49	-1.27	0.78891958
SOCS3	9021	227697_at	118	26	2.21	0.81490784
EPAS1	2034	200878_at	429	157	2.31	0.8417473

TRGC2	6967	216920_s_at	466	410	1.12	0.8773761
FOXP3	50943	221333_at	3	3	-1	0.93339042
CDKN1A	1026	202284_s_at	173	182	-1.04	0.94107376
TLR2	7097	204924_at	64	72	-1.07	0.96066244
GADD45G	10912	204121_at	11	11	1	0.98495784

Appendix A1.6 Genes transcriptionally activated by MITF

Gene Symbol	Gene Title	Entrez Gene ID
ACP5	acid phosphatase 5, tartrate resistant	54
BCL2	B-cell CLL/lymphoma 2	596
CDH1	cadherin 1, type 1, E-cadherin (epithelial)	999
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	1026
CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	1029
CLCN7	chloride channel 7	1186
CTSK	cathepsin K	1513
DCT	dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)	1638
MLANA	melan-A	2315
GZMB	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)	3002
HIF1A	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)	3091
ITGA4	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)	3676
MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)	4157
MET	met proto-oncogene (hepatocyte growth factor receptor)	4233
TRPM1	transient receptor potential cation channel, subfamily M, member 1	4308
NGFR	nerve growth factor receptor (TNFR superfamily, member 16)	4804
SERPINE1	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	5054
PRKCB	protein kinase C, beta	5579
PTGDS	prostaglandin D2 synthase 21kDa (brain)	5730
SILV	silver homolog (mouse)	6490
TBX2	T-box 2	6909
TPH1	tryptophan hydroxylase 1	7166
TPSAB1	tryptase alpha/beta 1	7177
TYR	tyrosinase (oculocutaneous albinism IA)	7299
TYRP1	tyrosinase-related protein 1	7306
BEST1	bestrophin 1	7439
CADM1	cell adhesion molecule 1	23705
PGDS	prostaglandin D2 synthase, hematopoietic	27306
OSTM1	osteopetrosis associated transmembrane protein 1	28962
OSCAR	osteoclast associated, immunoglobulin-like receptor	126014

Appendix A2 - CD