CELL INDEX DATABASE (CELLX): A WEB TOOL FOR CANCER PRECISION MEDICINE*

KEITH A. CHING[1], KAI WANG[1], ZHENGYAN KAN[1], JULIO FERNANDEZ[1], WENYAN ZHONG[1], JAREK KOSTROWICKI[1], TAO XIE[1], ZHOU ZHU[1], JEAN-FRANCOIS MARTINI[2], MARIA KOEHLER[2], KIM ARNDT[1], PAUL REJTO[1]

[1]Oncology Research Unit, [2]Oncology Business Unit, Pfizer Global Research & Development, Pfizer Inc., 10777 Science Center Drive San Diego, CA 92121, USA Email: keith.ching@pfizer.com

The Cell Index Database, (CELLX) (http://cellx.sourceforge.net) provides a computational framework for integrating expression, copy number variation, mutation, compound activity, and meta data from cancer cells. CELLX provides the computational biologist a quick way to perform routine analyses as well as the means to rapidly integrate data for offline analysis. Data is accessible through a web interface which utilizes R to generate plots and perform clustering, correlations, and statistical tests for associations within and between data types for ~20,000 samples from TCGA, CCLE, Sanger, GSK, GEO, GTEx, and other public sources. We show how CELLX supports precision oncology through indications discovery, biomarker evaluation, and cell line screening analysis.

1. Introduction

To support precision medicine patient selection strategies, genomics data is used to identify oncogenic drivers or dysregulated pathways in cancer cells susceptible to therapeutic intervention. Notably, efforts by The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov), the Cancer Cell Line Encyclopedia (CCLE)[1], and Sanger Wellcome Trust Genomics of Drug Sensitivity in Cancer (GDSC)[2] have generated a plethora of data and datatypes that can be used for generating patient selection hypotheses. However, multiple genomics data types such as expression, copy number variation (CNV), and mutation are large and unwieldy to manage. For the computational biologist, much time and effort can be spent to assemble an up to date table of features which can be computed on because new data are often generated frequently and incrementally. Thus, there is a need for an infrastructure to perform simple, quick, and routine analyses on multi-dimensional genomics data as well as the automated assembly of data tables for offline computation using more sophisticated algorithms.

Currently, there exist several cancer genomics databases to access expression, CNV, mutation, and integrated data as reviewed in [3]. For example, BioGPS[4] provides expression data, Tumorscape[5] contains CNV measurements, the Sanger Catalog of Somatic Mutations in Cancer (COSMIC)[6] lists mutations, and the cBio Portal[7] integrates multiple TCGA data types. Additionally, databases with compound activity data include GDSC and CCLE. Here we present a publicly available web-based informatics tool to integrate data, perform analysis, and visualize results from public as well as private internal sources to support precision medicine activities.

* This work is supported by Pfizer, Inc.

2. Architecture

The underlying MySQL database consists of 22 tables for expression, CNV, mutation, compound, sample, meta data, RNAi, RPPA, and gene annotation data. The Perl CELLX application runs on an Apache web server. R-serve (http://www.rforge.net/Rserve/) instances generate plots and perform statistical analyses. An Apache Tomcat application server runs a custom Java servlet which bridges Perl and R by funneling Perl http requests to the R-serves and sends results back to the web server. A demo site, instructions, source code, database dumps, and data parsing / loading scripts are available at http://cellx.sourceforge.net.

3. Gene Based Search

A common starting point for indications discovery is asking where the target of interest is altered. CELLX can plot the relative expression or CNV of a gene within a dataset or across multiple compatible datasets. For instance, RNA-Seq data processed by RSEM[8] can be compared across tumors profiled not only by TCGA, but CCLE as well. CDK4 expression can be seen to have high outliers in Glioblastoma Multiforme (GBM), melanoma (SKCM), breast (BRCA), Lower Grade Glioma (LGG), and sarcomas (SARC) (Figure 1). A similar plot can be generated of CNV to identify datasets with amplifications or deletions. CELLX can chart the relationship between expression and CNV across datasets using scatter plots of expression versus CNV. A hallmark of amplification, CDK4 expression levels scale with CNV level in several datasets (Figure 2a,b).

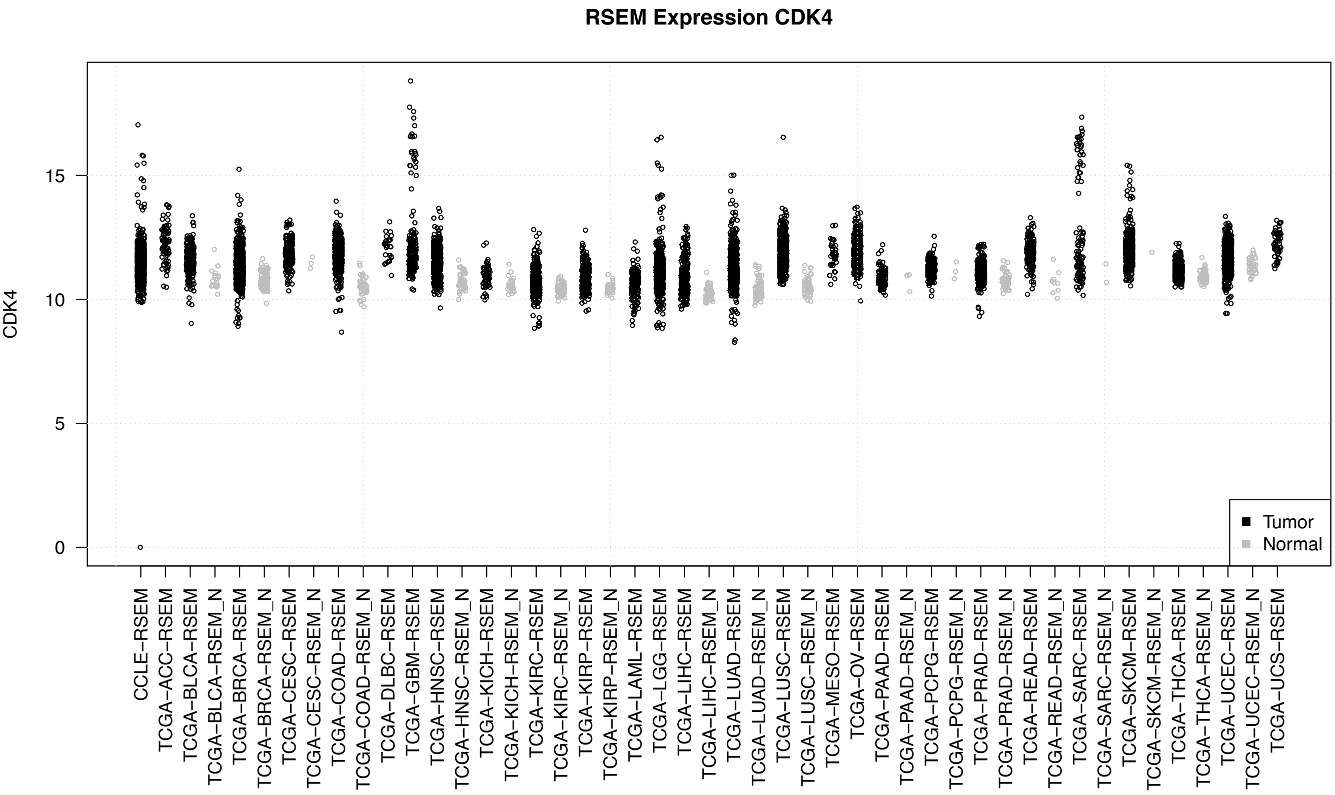


Figure 1. RNA-Seq RSEM gene expression of CDK4 (y-axis, log2) across datasets shows higher expression in tumor vs. adjacent normal tissue. Particular groups of outliers can be seen in GBM (glioblastoma multiforme), SARC (sarcoma), SKCM (skin cutaneous melanoma), LGG (brain lower grade glioma), and cell lines (CCLE).

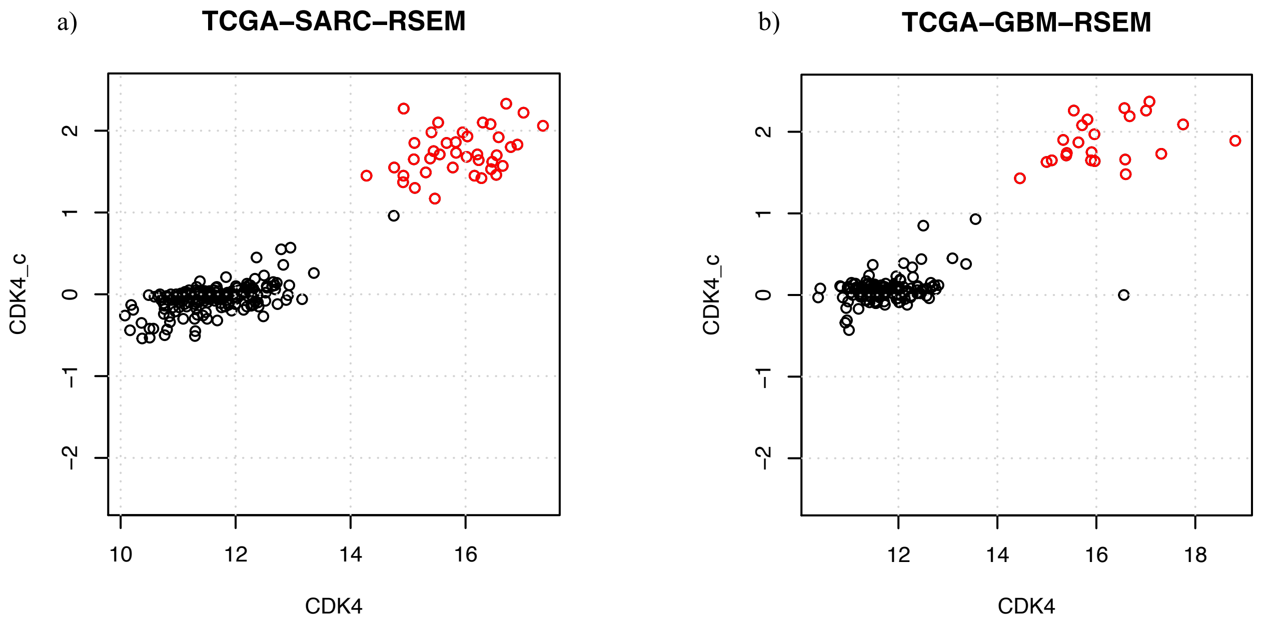


Figure 2. Correlation of expression and CNV. CNV (y-axis in log2 diploid genome) vs. RSEM expression levels (log2) for CDK4 show that a) SARC and b) GBM datasets have a sizable population of cells overexpressing CDK4 due to amplification of the locus. Additionally, expression levels scale with CNV levels. Clear outliers from the main distribution of CNV values can help determine appropriate CNV cut offs for amplification status. In this example, samples colored red have ≥ 1 log2 diploid genomes (i.e. ≥~4 copies).

4. Integrated Visualization

Mixed data types can be visualized in 2D scatter plots to look at the relationship between two datatypes on the same or different genes. For instance, expression of gene A on the x-axis can be plotted versus the CNV of gene B on the y-axis. Other plottable datatypes are protein levels for Reverse Phase Protein Arrays (RPPA), the mutation count per sample, the general amount of CNV per sample, IC50 values for compounds, and meta data. Multiple layers of data can be added to the plot to increase dimensionality. As a simple example, one can plot the expression of ERBB2 expression vs. ERBB2 CNV overlaid with ERBB2 mutations (Figure 3a) or breast cancer subtype meta data. (Figure 3b). The underlying data used to generate each plot is linked as a tab separated tsv file for downloading.

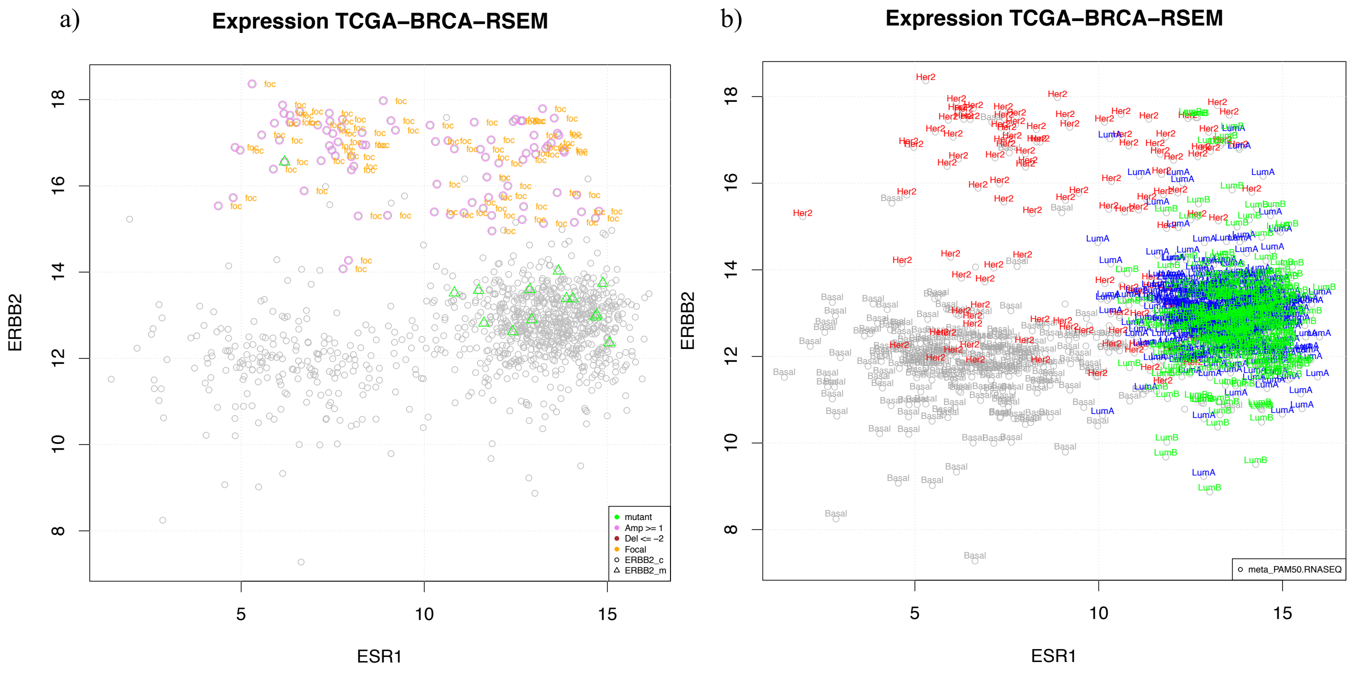


Figure 3. 2D scatter plots. a) Gene expression of ESR1 (x-axis, log2) vs. ERBB2 (y-axis, log2) gene expression. ERBB2 CNV over the selected threshold of 1 (log2 diploid genome) is colored pink. Focal amplifications (≤ 10MB) are denoted with ‘foc’. Mutations in ERBB2 are colored green. c) Meta data for PAM50 subtype classification are colored and overlaid on the ESR1 vs. ERBB2 gene expression plot.

5. Biomarker Frequency Reports

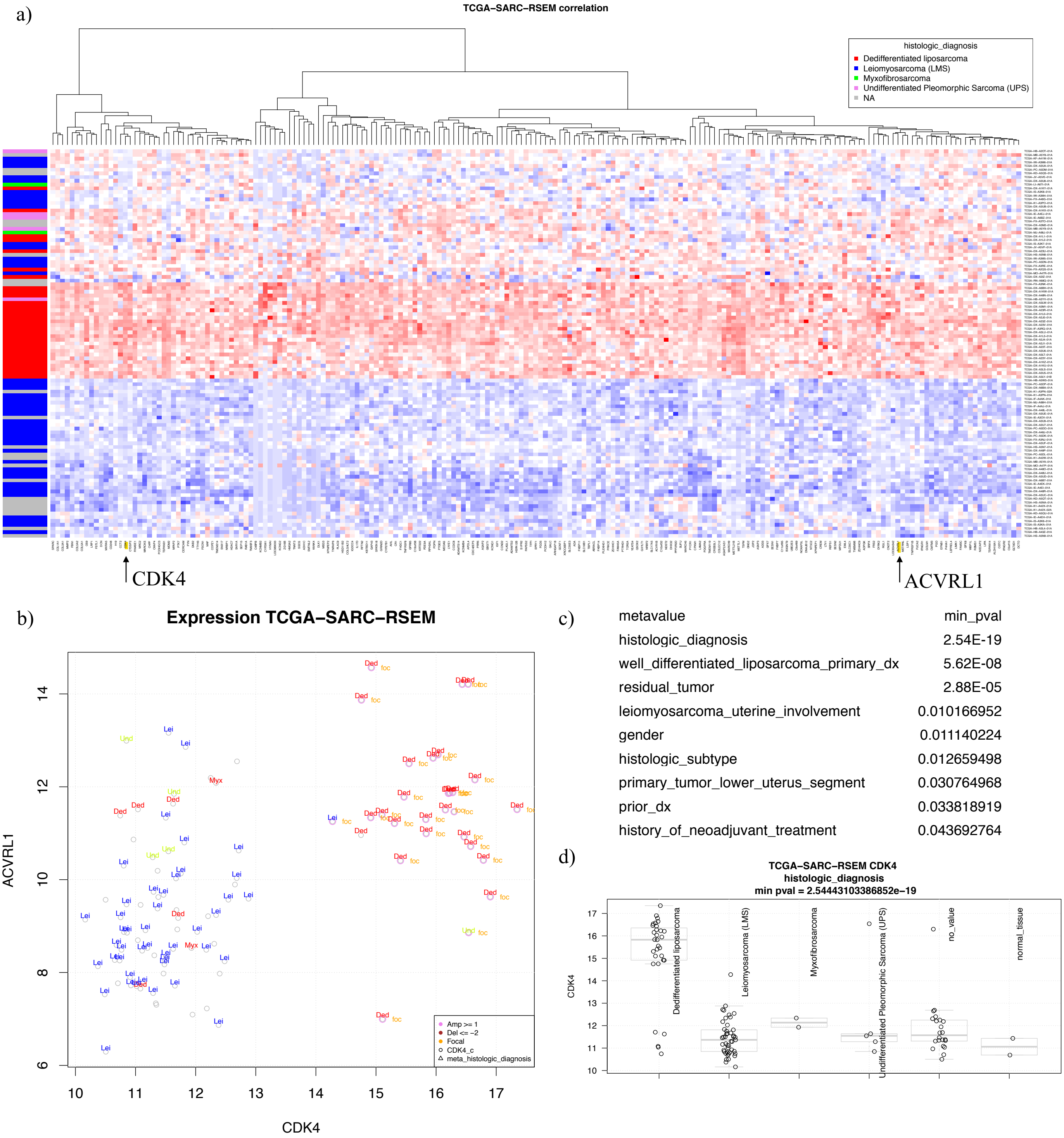
Tables of the frequency of alterations across datasets can help to prioritize indications for therapies with known biomarkers. For instance, the venn report of the frequency of CDK4 biomarker alterations within datasets shows significant frequencies of CDK4 amplification in sarcoma, gliomas, and melanoma TCGA datasets (Table 1). Cutoffs can be defined by expression level, CNV level, and/or mutation status. The co-occurrence or exclusion of 2-4 biomarkers within the same sample can also be quantified.

Table 1. Frequency report for CDK4 alterations in TCGA. CDK4_c is the number of samples in which the CNV exceeds the set threshold, in this case ~4 copies. CDK4_m is the number of samples with a CDK4 mutation. The cells_c/_m columns are the number of samples for which CNV or mutation data are available, respectively. Percentages are calculated as altered / total for each individual alteration type.

sourcename	CDK4_c	cells_c	CDK4_m	cells_m	cell_type	tumor_type	CNV%	MUT%
TCGA-SARC	35	171	0	0	soft_tissue	Sarcoma	20.47	NA
TCGA-GBM	73	607	0	150	neuronal	Glioblastoma multiforme	12.03	0
TCGA-LGG	14	471	1	612	neuronal	Brain Lower Grade Glioma	2.97	0.16
TCGA-ACC	2	90	0	91	adrenal_gland	Adrenocortical carcinoma	2.22	0
TCGA-SKCM	7	387	8	372	skin	Skin Cutaneous Melanoma	1.81	2.15
TCGA-LUAD	5	510	3	491	lung	Lung adenocarcinoma	0.98	0.61
TCGA-STAD	2	403	1	373	stomach	Stomach adenocarcinoma	0.5	0.27
TCGA-BRCA	5	1074	1	777	breast	Breast invasive carcinoma	0.47	0.13
TCGA-BLCA	1	255	2	242	urinary_tract	Bladder Urothelial Carcinoma	0.39	0.83
TCGA-OV	2	569	0	476	ovary	Ovarian serous cystadenocarcinoma	0.35	0
TCGA-LUSC	1	487	0	233	lung	Lung squamous cell carcinoma	0.21	0
TCGA-COAD	0	446	2	219	large_intestine	Colon adenocarcinoma	0	0.91
TCGA-PRAD	0	381	0	300	prostate	Prostate adenocarcinoma	0	0
TCGA-THCA	0	508	0	428	thyroid	Thyroid carcinoma	0	0
TCGA-PAAD	0	92	1	91	pancreas	Pancreatic adenocarcinoma	0	1.1
TCGA-PCPG	0	175	0	0	adrenal_gland	Pheochromocytoma and Paraganglioma	0	NA
TCGA-MESO	0	37	0	0	pleura	Mesothelioma	0	NA
TCGA-READ	0	164	0	1	rectum	Rectum adenocarcinoma	0	0
TCGA-UCEC	0	533	5	248	endometrium	Uterine Corpus Endometrial Carcinoma	0	2.02
TCGA-KIRC	0	521	6	328	kidney	Kidney renal clear cell carcinoma	0	1.83
TCGA-ESCA	0	126	0	0	oesophagus	Esophageal carcinoma	0	NA
TCGA-DLBC	0	28	0	79	haematopoietic_	Lymphoid Neoplasm Diffuse Large B-cell	0	0
TCGA-KICH	0	66	0	66	kidney	Kidney Chromophobe	0	0
TCGA-UCS	0	57	0	57	uterus	Uterine Carcinosarcoma	0	0
TCGA-KIRP	0	212	0	169	kidney	Kidney renal papillary cell carcinoma	0	0
TCGA-LAML	0	194	0	118	haematopoietic_	Acute Myeloid Leukemia	0	0
TCGA-LIHC	0	213	5	202	liver	Liver hepatocellular carcinoma	0	2.48
TCGA-HNSC	0	516	5	513	upper_aerodiges	Head and Neck squamous cell carcinoma	0	0.97
TCGA-CESC	0	206	0	41	cervix	Cervical squamous cell carcinoma and	0	0

6. Analysis

CELLX can identify genes whose expression correlates with a gene of interest and return a table of significant genes that can be visualized via a heat map with labelled metadata. For example, a search for genes correlated with CDK4 expression in the TCGA sarcoma dataset yields ACVRL1 which is expressed by vascular endothelium and a potential anti-angiogenesis target. (Figure 4a)



metavalue	min_pval
histologic_diagnosis	2.54E-19
well_differentiated_liposarcoma_primary_dx	5.62E-08
residual_tumor	2.88E-05
leiomyosarcoma_uterine_involvement	0.010166952
gender	0.011140224
histologic_subtype	0.012659498
primary_tumor_lower_uterus_segment	0.030764968
prior_dx	0.033818919
history_of_neoadjuvant_treatment	0.043692764

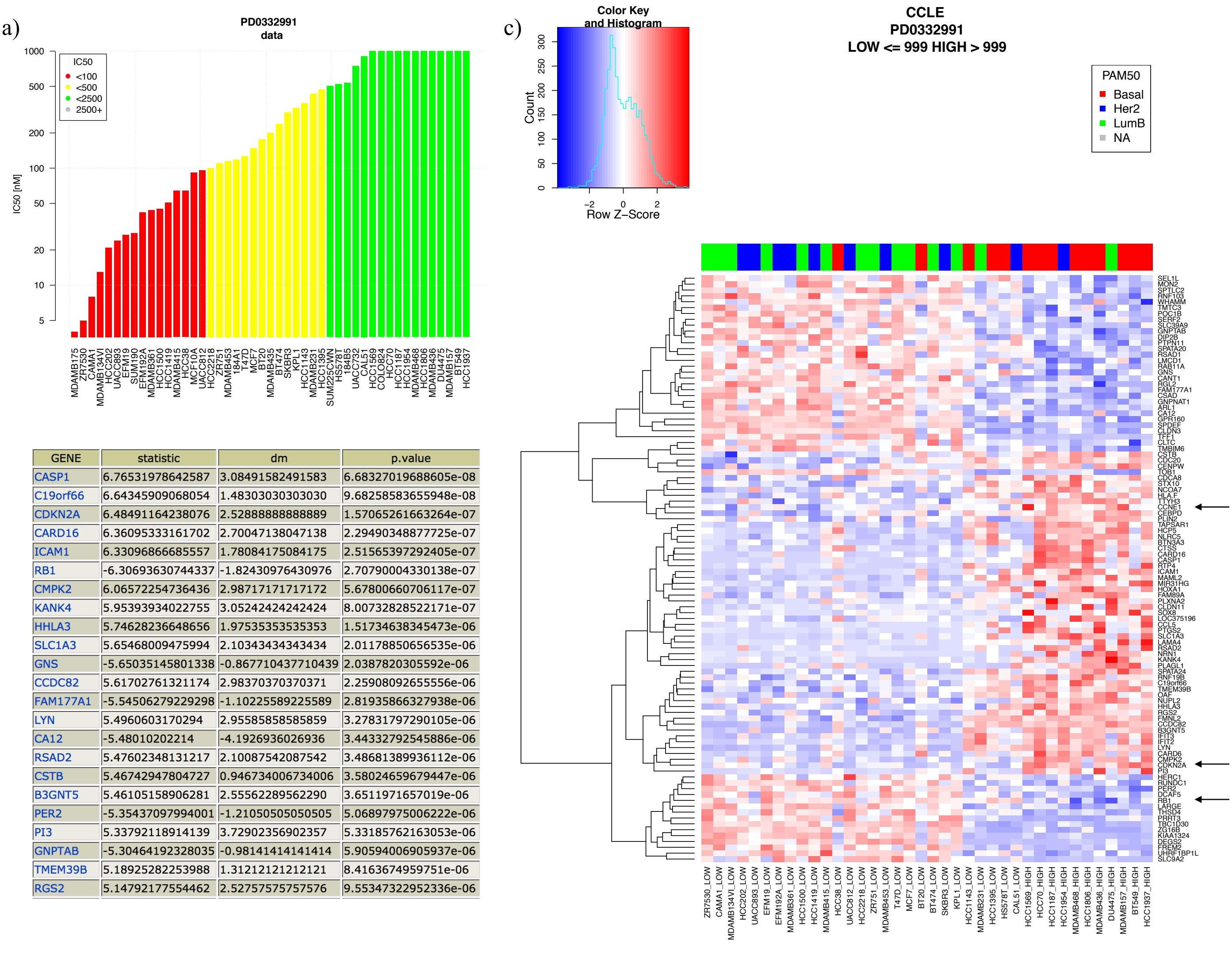
Figure 4. Analysis of features associated with CDK4 expression. a) Heatmap of top 200 genes (columns) correlated with CDK4 expression levels in samples (rows) from the TCGA sarcoma dataset showing ACVRL1 expression correlates with CDK4 (arrows). Meta data labels for histologic diagnosis are colored in a column on the left side of the plot. b) Scatter plot of CDK4 expression versus ACVRL1 expression showing high ACVRL1 expression in dedifferentiated liposarcomas. Metavalues from a) are colored and abbreviated by the first 3 letters. Amplification of CDK4 is denoted by a violet circle. foc=focal. c) Meta data with significantly different CDK4 expression levels. Min p-value is the lowest pairwise t-test score. d) Boxplot of histologic diagnosis by CDK4 expression data used in c).

A scatter plot of CDK4 vs. ACVRL1 shows higher ACVRL1 in Dedifferentiated Liposarcomas (DDPLS) vs. Leiomyosarcomas (Figure 4b). This is consistent with a study reporting immature and intermediate blood vessels in sarcomas and quantifying tumor microvessel density that is ~3X higher in DDLPS vs. Leiomyosarcomas. [9] The plot also shows that CDK4 expression is high in DDLPS and often focally amplified which is consistent with the literature.[10] CELLX can also test for significant gene expression associated with meta data features by performing a t-test of a gene's expression grouped by a sample's meta data. As an example, a search for meta data with significantly different CDK4 expression in the TCGA sarcoma dataset reveals that the histologic diagnosis type has large differences in CDK4 expression levels (lowest p-val = $2.54e^{-19}$) as calculated by a pairwise t-test between all groups (Figure 4c). A box plot of the groups from histologic diagnosis shows that the CDK4 values from DDPLS are higher than other sarcomas (Figure 4d). Additional types of analyses include the identification of differentially expressed genes using t-tests of gene expression between groups defined by a gene's expression, a gene's mutation status, or a meta value label. For example, one could ask what genes are differentially expressed between samples with high CDK4 vs. low CDK4, samples with mutated EGFR vs. wild type EGFR, or samples annotated as male vs. female. Conversely, one can search for mutated genes which differentially express the query gene. e.g. which gene(s) mutations have higher or lower expression of EGFR than wild-type.

7. Precision Medicine

To support precision medicine, CELLX can be used to generate responder / non-responder hypotheses from cell line screening data. As a retrospective example, one can analyze the cell line sensitivity profile of Palbociclib, a CDK4/6 inhibitor under development for ER+ breast cancer. Published breast cell line IC50 values for Palbociclib[11] show a range of responses. (Figure 5a) CELLX can associate IC50 values with cell line expression, CNV, and mutation data from data sources such as CCLE. Samples divided into two groups by user defined cutoffs, in this case <1uM for responder cell lines (LOW IC50) and ≥ 1uM for non-responder cell lines (HIGH IC50) can be used to identify genes whose expression is significantly different between responder and non-responder cells by calculating t-tests on the expression of ~20,000 genes and displaying a p-value ranked table (Figure 5b). Hierarchical clustering on the top 100 most significant genes, ordering the samples from low to high IC50, and coloring the samples by intrinsic breast subtype as defined by PAM50[12] shows that luminal B and Her2 subtypes tend to be sensitive to Palbociclib whereas cells of the basal subtype tend to be resistant (Figure 5c). Luminal A cell line subtypes were not represented in the screening set. Additionally, CELLX can dynamically generate a combination CNV / mutation table for genes which meet user defined amplification / deletion thresholds or have annotated mutations. A ranked table of p-values from Fisher's exact test for all genes with either a CNV or mutation alteration (Table 2) highlights genes potentially associated with compound activity. While individually, the appearance of any one gene is not necessarily significant, together the combined results from the expression, CNV, and mutation associations highlight RB1, CCNE1, and to a lesser extent CDKN2A. Specifically, the expression of RB1 was low in resistant cells whereas CDKN2A and CCNE1 were high in resistant cells. Interestingly, unlike other targeted therapies where the small molecule target is often the biomarker of sensitivity (e.g. EGFR, MET, BRAF) the significant Palbociclib biomarkers represent markers of resistance. RB1 deficiency (CNV deletion, STOP mutations, and low expression) and concomitant high CDKN2A expression[13] are characteristics of the basal or triple negative breast subtype status (Figure 5c). Thus, if most of the RB1 deficient samples belong to the triple negative subtype, the remaining luminal A/B (ER+/ERBB2+/-) and ERBB2+ segments would be enriched for possible CDK4i responders. In support of this notion, luminal B and Her2 breast subtype cell lines are mostly sensitive to CDK4i (Figure 5c).

CELLX can also confirm if the low RB1 expression found in triple negative breast cell lines also occurs in primary tissues by using the TCGA-BRCA breast invasive carcinoma dataset. CELLX can identify the genes that are most differentially expressed between RB1 high (≥ 9.5) vs. RB1 low (< 9.5) expressing cells using t-tests. Several of the top 100 ranking genes by p-value are related to cell cycle (RB1, CDKN2A, CCNE1) or DNA replication/repair (RFC2, RFC4, MCM5, MCM7, CDT1, NASP, POLK, POLD1, MUTYH, FANCE). Hierarchical clustering and labeling with the intrinsic subtype via PAM50[12] shows that similar to cell lines, we find that tumors with low RB1 and high CCNE1/CDKN2A expression are often of the basal subtype (Figure 6).



GENE	statistic	dm	p.value
CASP1	6.76531978642587	3.08491582491583	6.68327019688605e-08
C19orf66	6.64345909068054	1.48303030303030	9.68258583655948e-08
CDKN2A	6.48491164238076	2.52888888888889	1.57065261663264e-07
CARD16	6.36095333161702	2.70047138047138	2.29490348877725e-07
ICAM1	6.33096866685557	1.78084175084175	2.51565397292405e-07
RB1	-6.30693630744337	-1.82430976430976	2.70790004330138e-07
CMPK2	6.06572254736436	2.98717171717172	5.67800660706117e-07
KANK4	5.95393934022755	3.05242424242424	8.00732828522171e-07
HHLA3	5.74628236648656	1.97535353535353	1.51734638345473e-06
SLC1A3	5.65468009475994	2.10343434343434	2.01178850656535e-06
GNS	-5.65035145801338	-0.867710437710439	2.0387820305592e-06
CCDC82	5.61702761321174	2.98370370370371	2.25908092655556e-06
FAM177A1	-5.54506279229298	-1.10225589225589	2.81935866327938e-06
LYN	5.4960603170294	2.95585858585859	3.27831797290105e-06
CA12	-5.48010202214	-4.1926936026936	3.44332792545886e-06
RSAD2	5.47602348131217	2.10087542087542	3.48681389936112e-06
CSTB	5.46742947804727	0.946734006734006	3.58024659679447e-06
B3GNT5	5.46105158906281	2.55562289562290	3.6511971657019e-06
PER2	-5.35437097994001	-1.21050505050505	5.06897975006222e-06
PI3	5.33792118914139	3.72902356902357	5.33185762163053e-06
GNPTAB	-5.30464192328035	-0.98141414141414	5.90594006905937e-06
TMEM39B	5.18925282253988	1.31212121212121	8.4163674959751e-06
RGS2	5.14792177554462	2.52757575757576	9.55347322952336e-06

Figure 5. a) Waterfall plot of breast cell line responses to Palbociclib (PD0332991) colored by IC50 range. b) Example output listing the p-value of genes. dm = difference in group means, statistic = t-statistic (LOW-HIGH), p.value = uncorrected p-value of two-sided, two-class t-test with equal variances. Not shown: FDR and Hochberg adjusted p-values. c) Heatmap of gene expression of top 100 genes by t-test between sensitive (IC50 ≤ 999nM, LOW) and resistant cell lines (IC50 > 999nM, HIGH). The positions of RB1, CDKN2A, and CCNE1 are denoted with arrows. Cell lines are ordered by IC50 and colored by intrinsic breast subtype via PAM50.

Table 2. Association of mutations / CNV with response to Palbociclib (PD0332991). a) Ranking of genes by p-value for Fisher's Exact test. b) Breast cell line table of selected alterations. Breast cell lines are labeled LOW (sensitive) or HIGH (resistant) and marked altered or non-altered for mutation or CNV change in each gene. Cell lines are ordered by Palbociclib IC50 value. Genes with CNV values ≥ abs(1) or mutations from CCLE are marked as altered. CNV units are in log2 diploid genomes. (i.e. 1=~ 4 copies) CCLE mutation nomenclature: del = deletion, p.0 = whole gene deletion, ? = unknown change, fs = frameshift, * = STOP codon

a)

GENE	pval	GENE	pval
RB1	0.0004	ATP9B	0.0611
PIK3C2G	0.0048	CAPRIN1	0.0611
C19orf12	0.0136	CTIF	0.0611
CCNE1	0.0136	DNM2	0.0611
LOC284395	0.0136	EHF	0.0611
PLEKHF1	0.0136	ELP2	0.0611
POP4	0.0136	EPG5	0.0611
URI1	0.0136	FANCI	0.0611
VSTM2B	0.0136	HDLBP	0.0611
DOCK3	0.0136	LRP6	0.0611
NCOA4	0.0136	MAPK4	0.0611
ADRA1A	0.0136	MCPH1	0.0611
CTNNA1	0.0136	NKX6.3	0.0611
TCF12	0.0136	PDCD6	0.0611
CDH1	0.0459	PEBP4	0.0611
ANKS1B	0.0459	PTK2B	0.0611
DIP2C	0.0459	RP1L1	0.0611
GSTT1	0.0595	SGK223	0.0611
GSTTP2	0.0595	SMAD4	0.0611
LOC391322	0.0595	ZFYVE26	0.0611
D2HGDH	0.0611	MTAP	0.0932
DHRS4L1	0.0611	USP32	0.0932
DHRS4L2	0.0611	BCAS1	0.0932
ELAC1	0.0611	TRIM37	0.0932
GAL3ST2	0.0611	PIK3CA	0.0952
LINC00906	0.0611	TP53	0.0952
LINC01029	0.0611	AUTS2	0.0971
LOC100420587	0.0611	LOC649352	0.0971
LOC100505835	0.0611	MIR4650.1	0.0971
LOC102724958	0.0611	MIR4650.2	0.0971
LOC439994	0.0611	SIGLEC14	0.0971
MIR6511B1	0.0611	FHIT	0.0971
NAALADL2	0.0611	PIK3C2B	0.0971
NUTM2A.AS1	0.0611	PTEN	0.1176
RBFOX1	0.0611	CDKN2A	0.1362
SALL3	0.0611	LOC284344	0.1560
UGT2B28	0.0611	LPAR6	0.1560
UQCRFS1	0.0611	NRG1	0.1560
APC	0.0611	PDE4D	0.1560
BTK	0.0611	EEF2K	0.1560
ELN	0.0611	EPHB3	0.1560
EPHB6	0.0611	ITPR1	0.1560
GCNT2	0.0611	KIAA1549	0.1560
HIPK2	0.0611	MAP3K19	0.1560
KLK15	0.0611	MELK	0.1560
NOS2	0.0611	MLKL	0.1560
OMG	0.0611	MMP8	0.1560
TBX22	0.0611	MYLK	0.1560
ZNF142	0.0611	PLCB2	0.1560
AGPAT5	0.0611	SPTA1	0.1560

b)

cell_name	PD0332991	RESPONSE	RB1	PIK3C2G	CCNE1	CDKN2A
pvalue			0.0004	0.0048	0.0136	0.1362
MDAMB175	4	LOW				
ZR7530	5	LOW		p.P129del		
CAMA1	8	LOW				
MDAMB134VI	13	LOW		p.P129del		
HCC202	21	LOW				
UACC893	24	LOW				
EFM19	27	LOW				p.0?/-2.16
SUM190	28	LOW				
EFM192A	42	LOW				
MDAMB361	44	LOW		p.P129del		p.M52I
HCC1500	45	LOW		1.27		-2.24
HCC1419	51	LOW		p.P129del		
HCC38	64	LOW		p.P129del		p.0?/-2.75
MDAMB415	64	LOW		p.P129del		
MCF10A	92	LOW				
UACC812	96	LOW	1.26	p.P129del		
HCC2218	100	LOW		p.P129del		
ZR751	110	LOW		p.P129del		
MDAMB453	115	LOW		p.P129del		
184A1	118	LOW				
T47D	127	LOW		p.P129del		
MCF7	148	LOW				p.0?/-2.19
BT20	177	LOW	p.I388S	p.P129del		p.0?/-2.11
MDAMB435	201	LOW				p.?
BT474	240	LOW				-1.07
SKBR3	300	LOW				
KPL1	327	LOW				-1.97
HCC1143	359	LOW				
MDAMB231	432	LOW		p.P129del		p.0?/-2.53
HCC1395	472	LOW				p.0?/-2.03
SUM225CWN	503	LOW				
HS578T	524	LOW				p.0?
184B5	538	LOW				
UACC732	744	LOW				
CAL51	905	LOW		p.P129del		
MDAMB468	1000	HIGH	p.?/-1.89			
MDAMB436	1000	HIGH	p.G203fs*9			
HCC1954	1000	HIGH				
HCC1937	1000	HIGH	p.T738_R775del38			
DU4475	1000	HIGH	p.0?/-1.92			
HCC1569	1000	HIGH			2.02	
HCC1187	1000	HIGH				
BT549	1000	HIGH	p.?/-2.22			
MDAMB157	1000	HIGH			1.01	
COLO824	1000	HIGH	p.?			
HCC70	1000	HIGH	p.N480del			
HCC1806	1000	HIGH			1.25	p.0?/-2.25

8. Summary

CELLX is an informatics infrastructure to manage multi-dimensional genomics datasets containing expression, copy number variation, mutation, and compound sensitivity information. A browser based web page enables an accessible way to visualize, analyze, and download the database data in a pre-formatted table suitable for offline computation. CELLX is presently

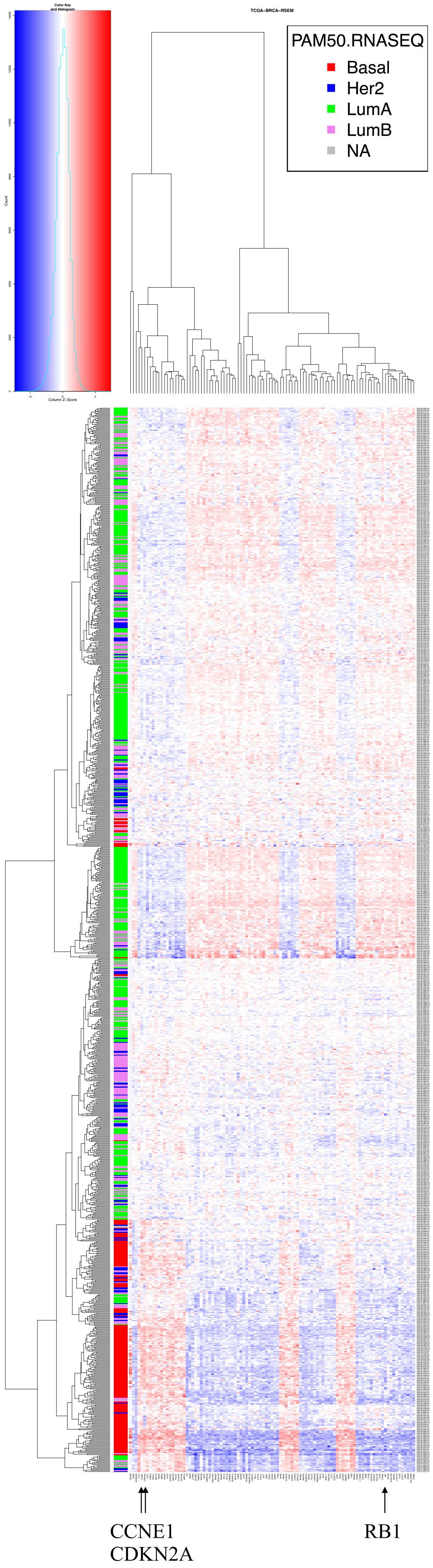


Figure 6. TCGA Breast differential gene expression between RB1 high and RB1 low expressing tumors. Hierarchical clustering of the top 100 genes in a heat map colored by breast subtype as determined by PAM50. Positions of CDKN2A, CCNE1, and RB1 are denoted by arrows.

focused on supporting oncology precision medicine through the evaluation of preconceived hypotheses as well as unbiased, data driven hypothesis generation. Though usable by the general user, CELLX is aimed at the computational biologist who desires more control over the data or wants to integrate custom data not available in public databases.

9. Data Processing

When available, summarized data from the source was used for TCGA, CCLE, and Tumorscape except for CNV calls. If Affymetrix SNP files were available, they were processed relative to the hg18 assembly using the aroma.affymetrix R package according to the methods of H. Bengtsson et al.[14] using the average baseline of 128 female HapMap samples[15] as the reference to maintain consistency and comparability across datasets. Microarray expression data from GEO, Sanger, and CCLE were GC Robust Multiarray Average normalized using R and the gcrma[16] library. Comparable to the TCGA RNA-Seq RSEM pipeline, CCLE RNA-Seq[17] data was processed using RSEM[8] on RefSeq sequences, quartile normalized to 1000, and log2 transformed. The R library genefu[18] predicted PAM50 subtypes and genefilter[19] enabled fast t-tests, F-tests, and correlations. Plots were made using CELLX and edited using Preview and Pages.

Acknowledgements

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/ dbGaP Study Accession: phs000178.v8.p7 We thank Andy Futreal and the Wellcome Trust Sanger Institute for generously providing access to cell molecular profiling data. We also thank Adam Pavlicek and Shibing Deng for help with R and Heather Estrella for discussion and feedback.

References

1. Barretina J, et.al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012 Mar 28;483(7391):603-7. PMID: 22460905
2. Yang W, et.al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. 2013 Jan;41:D955-61 PMID: 23180760
3. Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. *Genes Dev*. 2011 Mar 15;25(6):534-55. doi: 10.1101/gad.2017311. PMID:21406553
4. Wu C, Macleod I, Su AI. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res*. 2013 Jan;41:D561-5 PMID: 23175613
5. Beroukhim R, et.al. The landscape of somatic copy-number alteration across human cancers. *Nature*. 2010 Feb 18;463(7283):899-905. PMID: 20164920
6. Forbes SA, et.al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*. 2011 Jan;39:D945-50. PMID: 20952405
7. Cerami E, et.al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov*. 2012 May;2(5):401-4. PMID: 22588877
8. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011 Aug 4;12:323. doi: 10.1186/1471-2105-12-323. PMID:21816040
9. Baneth V, Raica M, Cîmpean AM. Rom J Assessment of angiogenesis in soft-tissue tumors. *Morphol Embryol.* 2005;46(4):323-7. PMID:16688371
10. Binh MB, Sastre-Garau X, Guillou L, de Pinieux G, Terrier P, Lagacé R, Aurias A, Hostein I, Coindre JM. MDM2 and CDK4 immunostainings are useful adjuncts in diagnosing well-differentiated and dedifferentiated liposarcoma subtypes: a comparative analysis of 559 soft tissue neoplasms with genetic data. *Am J Surg Pathol*. 2005 Oct;29(10):1340-7. PMID:16160477
11. Finn RS, Dering J, Conklin D, Kalous O, Cohen DJ, Desai AJ, Ginther C, Atefi M, Chen I, Fowst C, Los G, Slamon DJ. 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 Res*. 2009;11(5):R77. PMID:19874578
12. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009 Mar 10;27(8):1160-7. doi: 10.1200/JCO.2008.18.1370. Epub 2009 Feb 9. PMID:19204204
13. Knudsen ES, Knudsen KE. Tailoring to RB: tumour suppressor status and therapeutic response. *Nat Rev Cancer.* 2008 Sep;8(9):714-24. doi: 10.1038/nrc2401 PMID:19143056
14. Bengtsson H, Irizarry R, Carvalho B, Speed TP (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 24: 759– 767. PMID:18204055
15. HapMap data available from http://hapmap.ncbi.nlm.nih.gov/downloads/raw_data/hapmap3_affy6.0/ Originally obtained from Affymetrix, but no longer available from that source.
16. Wu J and Gentry RIwcfJMJ. gcrma: Background Adjustment Using Sequence Information. R package version 2.36.0. http://www.bioconductor.org/packages/release/bioc/html/gcrma.html
17. CCLE RNA-Seq data obtained from The Cancer Genomics Hub (CGHub) https://cghub.ucsc.edu/
18. Haibe-Kains B, Schroeder M, Bontempi G, Sotiriou C and Quackenbush J (2014). genefu: Relevant Functions for Gene Expression Analysis, Especially in Breast Cancer.. R package version 1.14.0, http://compbio.dfci.harvard.edu.
19. Gentleman R, Carey V, Huber W and Hahne F. genefilter: genefilter: methods for filtering genes from microarray experiments. R package version 1.46.1. http://www.bioconductor.org/packages/release/bioc/html/genefilter.html