Zhu *et al. BMC Genomics* (2016) 17:455
DOI 10.1186/s12864-016-2807-y

BMC Genomics

RESEARCH ARTICLE

Open Access

Outlier analysis of functional genomic profiles enriches for oncology targets and enables precision medicine

Zhou Zhu*, Nathan T. Ihle, Paul A. Rejto and Patrick P. Zarrinkar*

Abstract

Background: Genome-scale functional genomic screens across large cell line panels provide a rich resource for discovering tumor vulnerabilities that can lead to the next generation of targeted therapies. Their data analysis typically has focused on identifying genes whose knockdown enhances response in various pre-defined genetic contexts, which are limited by biological complexities as well as the incompleteness of our knowledge. We thus introduce a complementary data mining strategy to identify genes with exceptional sensitivity in subsets, or outlier groups, of cell lines, allowing an unbiased analysis without any *a priori* assumption about the underlying biology of dependency.

Results: Genes with outlier features are strongly and specifically enriched with those known to be associated with cancer and relevant biological processes, despite no *a priori* knowledge being used to drive the analysis. Identification of exceptional responders (outliers) may not lead only to new candidates for therapeutic intervention, but also tumor indications and response biomarkers for companion precision medicine strategies. Several tumor suppressors have an outlier sensitivity pattern, supporting and generalizing the notion that tumor suppressors can play context-dependent oncogenic roles.

Conclusions: The novel application of outlier analysis described here demonstrates a systematic and data-driven analytical strategy to decipher large-scale functional genomic data for oncology target and precision medicine discoveries.

Keywords: Outlier analysis, Functional genomics, Oncology, Cancer, Target identification, Precision medicine, Oncogene addiction, Synthetic lethality

Background

A major challenge in oncology drug discovery is the identification of tumor vulnerabilities that can lead to novel therapeutic targets, and linking these vulnerabilities to specific patient populations that are likely to benefit from pharmacological inhibition of these targets. While historically drug targets have originated from in-depth dissection of cancer biology, more recently tumor genome sequencing efforts such as The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov) and International Cancer Genome Consortium (ICGC) (https://icgc.org) have defined the genomic landscape and complexity for an ever growing number of tumor types and subtypes. However, with these approaches, it is becoming increasingly difficult to identify novel oncogenic drivers that are both pharmacologically accessible and applicable to a substantial number of patients.

Functional genomics offers an alternative means for target identification that is complementary to in-depth biology and sequencing. Gene silencing through sequence-specific targeting of mRNAs by RNA interference (RNAi) takes advantage of an endogenous cellular pathway [1, 2] and has become a powerful research tool by enabling high-throughput and systematic loss of function screens in cultured cells and model organisms [3]. One of the largest screens to date is Project Achilles at the Broad Institute, a pioneering effort that has utilized a lentivirally delivered short hairpin RNA (shRNA)

* Correspondence: Zhou.Zhu@pfizer.com; pzarrinkar@yahoo.com
Oncology Research Unit, Pfizer Worldwide Research & Development, La Jolla Laboratories, 10777 Science Center Drive, San Diego, CA 92121, USA

library to catalog the dependency of 216 cancer cell lines on 11,000 genes [4].

The analysis of genome-scale RNAi screens typically has focused on *a priori* partition of cell lines based on known biological or genetic contexts, such as the mutation of an established oncogene or tumor suppressor, followed by a comparison of the sensitivity patterns of the two groups to identify genes that, when knocked down, confer preferential sensitivity in one group over the other. This analytical approach has led, for example, to the discovery of *ARID1B* and *SMARCA2* as specific vulnerabilities for *ARID1A* and *SMARCA4*-mutant cancers, respectively [5, 6]. However, the need to pre-define the groups for interrogation represents an inherent limitation due to the incompleteness of our knowledge (e.g. granularity in functional consequence of genomic lesions) as well as biological complexities (e.g. the role of molecular and genetic contexts). Predefining groups, furthermore, calls for a separate analysis for each biological or genetic context, making it impractical to query all contexts of potential interest.

To address these limitations, we have developed a complementary data mining strategy based on patterns of sensitivity in functional genomics screens that requires no *a priori* assumptions about the underlying biology of dependency. Oncogene addiction or synthetic lethality usually results in exceptional response in a subset of tumors or cell lines that are exquisitely vulnerable to knockdown or inhibition of the gene being interrogated [7]. The responder subsets are, by definition, outliers relative to the rest of the population or cell line panel. Taking advantage of this observation, our strategy adapts and extends outlier analysis methodologies to identify genes with a subset of exceptional responders among the screened cell lines. Such a data-driven approach in principle makes it possible to identify vulnerabilities in any biological or genetic context in a single analysis, and also allows for the discovery of novel or complex contexts in which inhibition of specific genes represents a vulnerability that would not have been considered in a pre-defined class comparison analysis.

Outlier analysis has been widely applied to gene expression data for the discovery of cancer-associated genes [8]. It was first described in the identification of a gene fusion in prostate cancer involving two transcription factors, *ERG* and *ETV1* [9], which led to the Cancer Outlier Profile Analysis (COPA) method [9, 10]. Many technically more sophisticated approaches have followed, including model-based pattern recognition for deviation from uni-modality [11–14] and numerical detection for marked high expression in a subset of tumors that is distant from the majority [15–19]. Outlier detection has also been useful in finding drugs with rare but exceptional response in clinical trials [7]. While highly informative, exceptional responder studies in the clinic are constrained by the relatively modest number of biological mechanisms currently targeted by drugs as well as the challenge of following up hypotheses in patients. Large-scale functional genomic studies relieve these restrictions and enable investigating thousands of genes in parallel.

Here we apply an outlier analysis based strategy to functional genomic profiles for systematic oncology target discovery. The utility of such approach is illustrated by the observation that genes with outlier patterns are strongly and specifically enriched with those known to be associated with cancer and relevant biological processes, despite no molecular profiling or any other information being used to drive the analysis. We show that it may enable the identification of novel candidate therapeutic targets, and that the characteristics of the exceptional responder lines could further point to tumor indications and biomarkers of response to guide precision medicine strategies.

Results

Identification of genes with outlier sensitivity patterns

To identify genes with an exceptional responder pattern, we used the union of the results output by three diverse methods. They each focus on different features (bimodality, variability, gap) to detect outliers and therefore are considered complementary. Application of these approaches is not intended as a comprehensive comparison of various outlier methodologies; rather we reasoned that together they would provide a more complete set of outliers and outlier genes than any single algorithm. The first two methods were originally developed for outlier analysis of gene expression data: Profile Analysis using Clustering and Kurtosis (PACK) [13], and Outlier Sum (OS) [15]. PACK is a model-based pattern recognition algorithm for discovering bimodal distribution, which first determines the number of clusters in the dataset for each gene and then computes a measure of how much the distribution differs from Gaussianity (kurtosis) for those gene profiles with two clusters. Positive kurtosis indicates clusters of unequal relative size, while negative kurtosis indicates clusters of approximately equal representation. The OS algorithm uses the “outlier-sum” statistic, which is defined using values outside a variability-based numerical limit. It was recently assessed to have the best performance among six closely related outlier techniques [20]. For the third method, inspired by the Q (gap-to-range ratio) statistic utilized in Dixon’s test for outlier detection [21], we devised an intuitive nonparametric approach (details described in Methods; Additional file 1: Figure S1 and Additional file 2) to explicitly identify genes with dependency patterns where groups of sensitive cell lines are separated by a major "gap" from the bulk population (Gap Analysis Procedure, or 'GAP' in short).

All three methods were applied to the Achilles (v2.4.3) ATARiS gene level scores across a diverse panel of 216 cell lines derived from consensus clustering of individual hairpin patterns for on-target effects [22]. The Achilles dataset includes 5299 genes that yielded similarity solution(s) and thereby consensus scores [4]. The PACK algorithm identified 793 genes with bimodal distribution, of which 571 (72 %) had positive kurtosis with one of the two clusters representing a small "outlier" subgroup. As the outlier subgroup could have higher or lower shRNA score than the rest of the panel, we focused on the 105 genes for which the outlier group is more vulnerable to knockdown. The OS and GAP methods led to the identification of genes with a non-random outlier pattern (FDR ≤ 0.05) where statistical significance was estimated using permutation. We also required that the outlier group consist of at least five cell lines to avoid spurious one-off observations, yielding 90 genes from the PACK algorithm, 84 from the OS method, and 72 from the GAP approach. In all, there were 169 unique outlier genes, including 16 that emerged from all three methods (Fig. 1a and b).

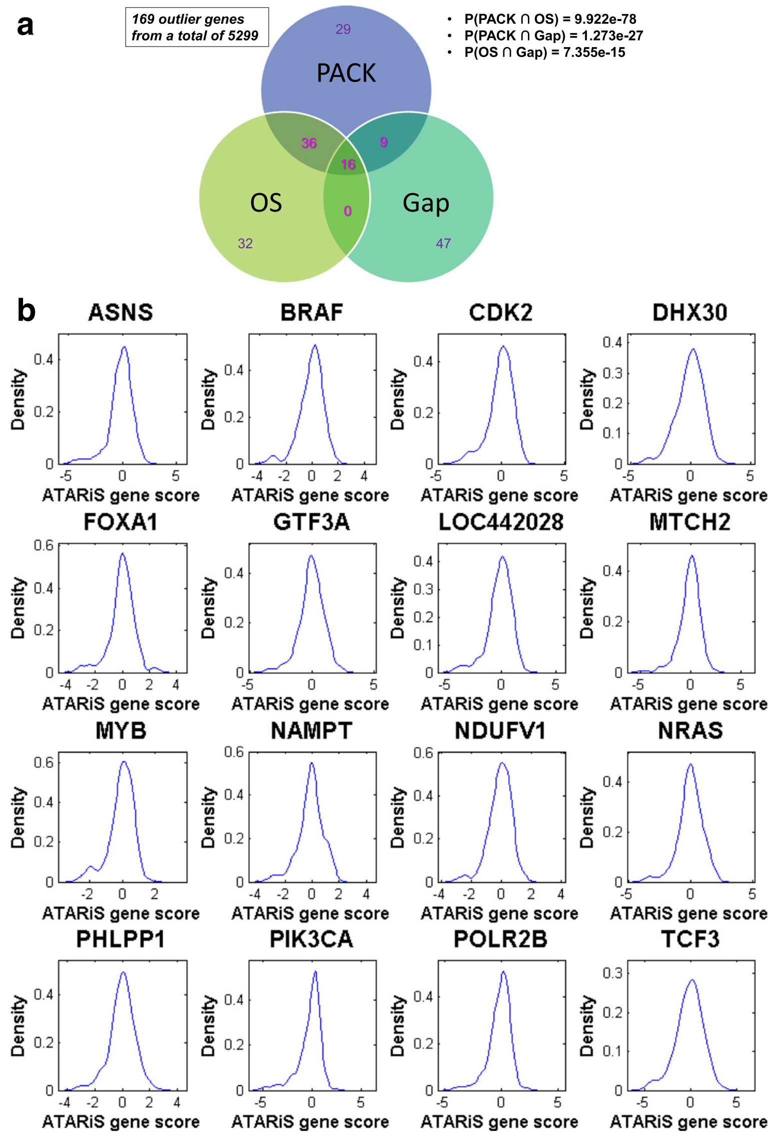

Fig. 1 Outlier genes identified with Profiling Analysis using Clustering & Kurtosis (PACK), Outlier Sum (OS) and Gap Analysis Procedure (GAP) methodologies. **a** A summary Venn diagram including statistical significance of pairwise overlap (determined using cumulative hypergeometric probability distribution), with detailed gene list and relevant results included in Additional file 4: Table S2. **b** The ATARiS gene level score distribution for the 16 genes identified by all three outlier methods. A probability density estimate is computed by Gaussian kernel smoothing

Enrichment for curated cancer genes and oncogenic pathways

We evaluated the biological relevance of the identified outlier genes by comparing them to two well-known cancer gene collections. The Cancer Gene Census (CGC) from the Sanger Institute catalogues genes for which mutations have been causally implicated in cancer [23] while the MSK-IMPACT™ (Integrated Mutation Profiling of Actionable Cancer Targets) panel is a curated collection of key cancer genes used for diagnostic genomic testing with next-generation sequencing technology (https://www.mskcc.org/msk-impact). Although determined independent of any genetic or other molecular profiling information, the genes with outlier patterns are significantly over-represented by those genetically linked to cancer from both CGC and IMPACT (Fig. 2). Each of the three outlier analysis approaches enriched for established cancer genes, and their union had greater statistical significance than any of them alone (Fig. 2), supporting the complementarity of these methodologies.

We further assessed the validity of our outlier analysis strategy through unbiased signature comparison. Among 1330 gene sets from the Broad Institute's MSigDB canonical pathway (CP) library (http://www.broadinstitute.org/gsea/msigdb), which collects canonical representations of biological processes compiled by domain experts from pathway databases, the outlier genes are strongly and specifically enriched in those associated with various tumor types including the prominent oncogenes and therapeutic targets *BRAF*, *NRAS*, *KRAS*, *PIK3CA*, *CDK4* and *CTNNB1*, as well as important cellular pathways and processes involved in cancer such as the Wnt, PI3K-mTOR, p53 and Rb-E2F pathways, cell cycle and apoptosis (Table 1). *CDK2* (Fig. 1b), like the targets of the recently approved breast cancer drug Ibrance, is a member of the cyclin-dependent kinase family of Ser/Thr protein kinases whose deregulation occurs frequently in certain types of cancer [24]. In addition to established cancer genes, our outlier analysis also reveals potentially novel therapeutic opportunities. For instance, asparagine synthetase (*ASNS*) (Fig. 1b), an enzyme that catalyzes the conversion of aspartate and glutamine to asparagine and glutamate in an ATP-dependent manner, has been shown to increase the chemotherapy sensitivity of leukemia cells resistant to L-asparaginase when inhibited [25, 26]. Collectively these results demonstrate that outlier analysis can serve as a useful strategy to identify cancer driver genes from RNAi sensitivity patterns.

Enabling tumor indication and predictive biomarker discoveries

With the identification of genes whose knockdown confers exceptional response, we next asked whether the cell lines making up the outlier groups are enriched in a particular tumor type or subtype, or have molecular features which may be linked to susceptibility to target interference. Such insights would not only guide the precision medicine strategy in oncology drug development for selecting patients who are most likely to benefit from targeted therapies, but also help to further prioritize the most biologically compelling and clinically translatable outlier patterns.

The 216 cell lines from the Achilles dataset were mapped to both general tumor types and common histological subtypes (Additional file 3: Table S1). In the case of breast cancer, both clinical marker (ER/PR/HER2 Triple Negative [TNBC], ER Positive [ER+], HER2 positive [HER2+]) and gene expression (Basal A [BaA_subtype], Basal B [BaB_subtype], Her2 [Her2_subtype], Luminal [Lum_subtype]) based classifications [27] were incorporated.

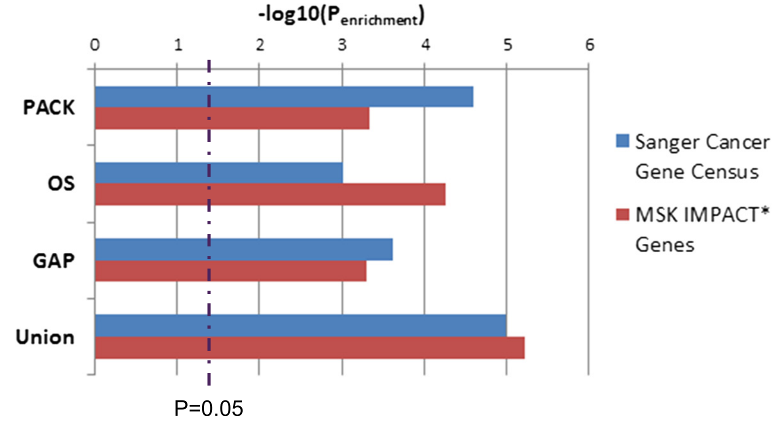

Fig. 2 Assessment of genes with significant outlier pattern through comparison with two well-known cancer gene collections. The Cancer Gene Census (CGC) from the Sanger Institute catalogues genes for which mutations have been causally implicated in cancer; The MSK-Integrated Mutation Profiling of Actionable Cancer Targets (IMPACT™) is a curated panel of key cancer genes used for diagnostic genomic testing. "Union" corresponds to 169 combined genes resulting from the three outlier methods (Fig. 1a)

Table 1 Enriched signatures from the Broad canonical pathway library among outlier genes

Canonical pathway signature	# of outlier genes in the signature	P(enrichment)	FDR
KEGG_PATHWAYS_IN_CANCER	17	1.53E-07	1.63E-04
KEGG_PROSTATE_CANCER	10	2.46E-07	1.63E-04
KEGG_THYROID_CANCER	6	5.01E-06	2.22E-03
PID_WNT_CANONICAL_PATHWAY	4	1.05E-05	3.25E-03
KEGG_ENDOMETRIAL_CANCER	7	1.22E-05	3.25E-03
PID_BETACATENIN_DEG_PATHWAY	4	3.65E-05	7.14E-03
KEGG_MELANOMA	7	3.76E-05	7.14E-03
KEGG_GLIOMA	7	4.44E-05	7.38E-03
REACTOME_CELL_CYCLE_MITOTIC	12	5.20E-05	7.68E-03
KEGG_BLADDER_CANCER	5	1.10E-04	1.47E-02
KEGG_NON_SMALL_CELL_LUNG_CANCER	6	1.56E-04	1.88E-02
PID_HES_HEYPATHWAY	5	1.76E-04	1.88E-02
KEGG_COLORECTAL_CANCER	6	1.84E-04	1.88E-02
REACTOME_SIGNALING_BY_FGFR	7	2.12E-04	2.02E-02
KEGG_ACUTE_MYELOID_LEUKEMIA	6	2.90E-04	2.41E-02
PID_SMAD2_3NUCLEARPATHWAY	6	3.35E-04	2.47E-02
ST_WNT_BETA_CATENIN_PATHWAY	4	3.53E-04	2.47E-02
BIOCARTA_GSK3_PATHWAY	4	3.53E-04	2.47E-02
REACTOME_MITOTIC_G2_G2_M_PHASES	5	3.90E-04	2.60E-02
REACTOME_SIGNALING_BY_FGFR_IN_DISEASE	7	4.24E-04	2.68E-02
PID_MTOR_4PATHWAY	6	5.70E-04	3.29E-02
BIOCARTA_G1_PATHWAY	4	5.94E-04	3.29E-02
SA_G1_AND_S_PHASES	3	6.95E-04	3.29E-02
REACTOME_G0_AND_EARLY_G1	3	6.95E-04	3.29E-02
PID_TCRRASPATHWAY	3	6.95E-04	3.29E-02
PID_RB_1PATHWAY	6	7.27E-04	3.29E-02
PID_IL4_2PATHWAY	5	7.61E-04	3.29E-02
REACTOME_NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE	7	7.77E-04	3.29E-02
KEGG_RENAL_CELL_CARCINOMA	6	8.17E-04	3.29E-02
REACTOME_DOWNSTREAM_SIGNALING_OF_ACTIVATED_FGFR	6	8.17E-04	3.29E-02
REACTOME_CELL_CYCLE	12	9.15E-04	3.54E-02
REACTOME_SIGNALING_BY_ERBB2	6	1.27E-03	4.21E-02
KEGG_CHRONIC_MYELOID_LEUKEMIA	6	1.27E-03	4.21E-02
BIOCARTA_CELLCYCLE_PATHWAY	3	1.44E-03	4.55E-02
PID_FOXM1PATHWAY	4	1.66E-03	4.96E-02
REACTOME_SIGNALING_BY_PDGF	6	1.89E-03	4.96E-02
REACTOME_PROLONGED_ERK_ACTIVATION_EVENTS	3	1.94E-03	4.96E-02
PID_E2F_PATHWAY	5	1.97E-03	4.96E-02
PID_ER_NONGENOMIC_PATHWAY	4	1.98E-03	4.96E-02
REACTOME_SIGNALING_BY_EGFR_IN_CANCER	6	2.07E-03	5.10E-02
BIOCARTA_P53_PATHWAY	3	2.55E-03	5.84E-02
KEGG_PANCREATIC_CANCER	5	2.77E-03	6.14E-02
PID_PI3KPLCTRKPATHWAY	4	3.64E-03	7.44E-02
PID_AR_TF_PATHWAY	4	3.64E-03	7.44E-02

Table 1 Enriched signatures from the Broad canonical pathway library among outlier genes *(Continued)*

BIOCARTA_IGF1R_PATHWAY	3	4.07E-03	8.07E-02
REACTOME_SIGNALING_BY_CONSTITUTIVELY_ACTIVE_EGFR	3	4.07E-03	8.07E-02
KEGG_ERBB_SIGNALING_PATHWAY	5	4.19E-03	8.07E-02
REACTOME_SIGNALING_BY_SCF_KIT	5	4.61E-03	8.63E-02
REACTOME_SIGNALING_BY_ERBB4	5	4.61E-03	8.63E-02
REACTOME_UNFOLDED_PROTEIN_RESPONSE	4	4.74E-03	8.76E-02
BIOCARTA_WNT_PATHWAY	3	6.04E-03	9.36E-02
BIOCARTA_BAD_PATHWAY	3	6.04E-03	9.36E-02
BIOCARTA_IGF1MTOR_PATHWAY	3	6.04E-03	9.36E-02
PID_AURORA_A_PATHWAY	3	6.04E-03	9.36E-02
WNT_SIGNALING	4	6.05E-03	9.36E-02
PID_AR_PATHWAY	4	6.79E-03	0.10
REACTOME_SIGNALLING_BY_NGF	7	7.50E-03	0.10

Significant tumor type associations (Additional file 4: Table S2) were found for more than half (86/169) of the genes with outlier patterns. These include well recognized relationships such as *KRAS* in pancreatic and colorectal cancers ($P = 1.969$e-10 and 2.180e-2 respectively), *CDK4* in luminal/ER+ breast cancer ($P = 3.968$e-7 and 9.044e-4 respectively), *BRAF* in skin cancer ($P = 3.109$e-3), *PIK3CA* in breast and gastric cancers ($P = 1.343$e-4 and 1.689e-3 respectively) and *APC* in colorectal cancer ($P = 2.581$e-4), as well as mechanistically supported relationships like *CDK2* in ovarian cancer ($P = 6.149$e-03; Table 2) where *CCNE1*, the cyclin that interacts with *CDK2*, is frequently amplified [28, 29].

To uncover predictive genetic biomarkers, we focused on the lesions that are most likely to be functional, as tumor genomes are often unstable and thus the vast majority of genetic changes are generally passengers [30]. For somatic mutations, we selected those in hotspot positions as well as nonsense and frameshift events. Hotspots were identified systematically using patient-derived genomic profiles from 20 TCGA tumor types (details described in Materials and Methods), which were subsequently employed to filter cell line mutation data compiled from the COSMIC [31] and CCLE [32] databases. For copy number alterations, we restricted our analysis to high-level amplifications (≥4 copies) and deletions (≤1 copy).

Six (*KRAS*, *NRAS*, *BRAF*, *PIK3CA*, *APC* and *ZMIZ1*) of the 169 genes with outlier patterns are significantly ($P < 0.05$) characterized by their own genetic lesions (Table 2), suggesting oncogenic addiction as the underlying mechanism of their exquisite vulnerability. All of these with the exception of *ZMIZ1* represent prominent oncogenes or tumor suppressors. *ZMIZ1* has previously been identified as a candidate oncogene in multiple murine transposon and insertional mutagenesis screens [33–35].

For 25 additional genes with outlier patterns, the associated predictive genetic biomarkers are either members of the same gene family or components of the same protein complex (Table 2), revealing possible synthetic lethal relationships where lesions in functionally related gene(s) confer special dependency. These include well-documented relationships such as the vulnerability for *CTNNB1* knockdown in the context of *APC* lesions and *E2F1* vulnerability in the context of *RB1* lesions as well as the recently discovered dependency on *ARID1B* in the context of *ARID1A* lesions [5]. In the case of *CDK2*, *CCNE1* amplification is specifically over-represented among its associated exceptional responder cell lines ($P = 8.215$e-04; Table 2), consistent with the tumor type enrichment of ovarian cancer described above. Novel relationships of potential interest were also observed such as *TOP2A* vulnerability with *PARP1* lesions, *HNRNPA1* vulnerability with *RPL22* lesions and *PSMD3* vulnerability for *PSMC4* lesions.

Distinct dependency and coherence between solid and hematological malignancies

To obtain a global view of tumor cell dependency, we performed unsupervised hierarchical clustering of functional genomic profiles using the genes with outlier dependency patterns (Additional file 4: Table S2). The vast majority of hematological cell lines cluster together by functional data (Additional file 5: Figure S2A and Additional file 6), in contrast to those of solid origin that are more scattered and heterogeneous, despite hematological cell lines representing only 14 % of the total, with tumor types like central nervous system (CNS) and ovarian cancer being equally or more heavily covered by the panel (Additional file 3: Table S1). Even functionally related genes, such as those in the Wnt pathway, differ in points of liability: while cell lines derived from liquid tumors tend to be

Table 2 Outlier genes whose predictive genetic biomarkers are the gene itself (in *italic*) or from the same gene family/protein complex, suggesting potential oncogenic addiction or a synthetic lethal relationship. Genetic biomarkers and tumor types are listed by decreasing statistical significance of association

Gene	Genetic biomarker (related)	Tumor type
APC	*APC*	*CRC*
ARID1B	ARID1A	
ATP5A1	ATP5H	
BRAF	*BRAF*	*CRC;Melanoma*
CDK2	CCNE1;SKP2	Ovarian;Ovarian_serous_adenocarcinoma;Breast Lum_subtype
CRNN	DST;SDF4	ALL;Leukemia/Lymphoma
CTNNB1	APC;PXN;CREBBP	CRC
DHX30	PARP1;DHX40	Breast;Breast HER2+;Breast Her2_subtype
E2F1	RB1	
FOXA1	FOXP4	Breast;Breast Lum_subtype;Breast HER2+;Prostate; Breast BaA_subtype;Prostate_adenocarcinoma; Breast Her2_subtype;Breast ER+
HNRNPA1	RPL22	
KRAS	*KRAS*	*Pancreatic;CRC*
LRPPRC	NFKBIB	Breast Lum_subtype;Breast;CRC; Ovarian_clear_cell_adenocarcinoma;Breast HER2+
NRAS	*NRAS*	
PIK3CA	*PIK3CA*	*Breast;Gastric;Breast Lum_subtype;Gastric_adenocarcinoma; CRC;Breast HER2+;Breast TNBC;Prostate_adenocarcinoma; Breast Her2_subtype;Breast ER+*
PSMD3	PSMC4;PSMD8	
RBBP4	ACTL6A;GNB4;WDR89	
RBM47	IGF2BP1;MAK16	Breast HER2+;Breast;Breast Her2_subtype; Esophageal_adenocarcinoma
RBMXL1	MAK16;TAF15	
RPS17	RPL38;RPL22	Ovarian_clear_cell_adenocarcinoma;Ovarian
RREB1	ZNF652;GLI2	Breast;Breast HER2+;Breast Her2_subtype
SF3A3	PRPF3	
SLC25A40	SLC2A3;SLCO1B1;SLCO1A2;SLC2A14;LST-3TM12	
SPEN	DHX38;CSTF2T	
TOP2A	PARP1;PPM1D;PRKDC	Breast;Breast Lum_subtype;Breast HER2+
TOPBP1	NBN;BRCA1	
TUBG1	APC	CRC
WDR18	WDR67;GNB2L1;WDR16;NWD1	
ZMIZ1	*ZMIZ1*	*Leukemia/Lymphoma;Breast;AML*
ZNF234	ZNF331	Multiple_Myeloma;Ewing_Sarcoma;Leukemia/Lymphoma;Bone
ZNF236	GLI1	

vulnerable to *TCF3* knockdown, those of solid origin are susceptible to other pathway genes including *GSK3A/B*, *CTNNB1* and *APC* (Additional file 5: Figure S2B).

Potentially context-dependent oncogenic roles for some tumor suppressor genes

Several outlier genes whose knockdown confers striking vulnerability are tumor suppressors including *APC* [36], *PHLPP1* [37] and *SPEN* [38], while a number of other candidates have been reported to harbor anti-oncogenic activities such as the pro-apoptotic gene *MTCH2* [39] and the anti-metastatic RNA chaperone *RBM47* [40]. Among the five lines most responsive to Adenomatous polyposis coli (*APC*) knockdown, all contain loss of function mutation in the *APC* gene itself and four are of colorectal (CRC) origin (Additional file 7: Figure S3A). Furthermore, they strongly overlap with those dependent on *CTNNB1* ($P = 1.219\text{e-}03$; Additional file 7: Figure

S3B). Truncating *APC* mutations have been reported to have a dominant negative effect on proliferation, spindle checkpoint control, survival and chromosomal stability [41]. The exceptional responders for *PHLPP1* and *MTCH2* (Fig. 1b) are over-represented by cells from leukemia/lymphoma and colorectal cancer respectively (*P* = 2.211E-8 and 5.178E-3; Additional file 3: Table S2). The results therefore support and may further generalize the notion that in certain contexts, tumor suppressor genes can become oncogenic and create specific liabilities.

Discussion

The analysis of large-scale cell based functional genomics datasets has predominantly focused on identifying vulnerabilities in known biological contexts, such as mutated oncogenes or tumor suppressors. A limitation of this approach is that it requires knowledge of the biology to be interrogated and can generally be described as starting with known biology, then looking for patterns in the data that support it. We have reversed this paradigm and now describe an alternative data mining strategy that starts by looking for profiles indicative of potential dependencies of interest, with no assumptions about the underlying biology of the dependency. It is based on identifying genes with subgroups of exceptionally sensitive cell lines. By definition, these exceptional responders are statistical outliers. We therefore hypothesized that outlier analysis, whose previous application in genomic studies has been limited to gene expression data, would also be useful in investigating RNA interference response patterns. We tested the hypothesis by applying outlier analysis to genome-scale shRNA screen results from Project Achilles at the Broad Institute, and found genes with outlier patterns are significantly and specifically enriched with those that have been causally or genetically linked to cancer as well as related pathways and processes, demonstrating the effectiveness of our novel approach.

As we intentionally selected three diverse methods for outlier analysis, it is not surprising that in addition to common predictions, each algorithm also identifies unique outlier genes (Fig. 1a). The significant over-representation of known cancer genes from each method (Fig. 2) suggests the general utility of the outlier approach is unlikely tied to a specific algorithm, and other outlier analysis methodologies can be similarly employed to decipher functional genomic profiles. The complementarity of the diverse approaches is manifested in the superior enrichment of their output union (Fig. 2), highlighting the heterogeneous pattern of outlier distribution among cancer genes.

The GAP method (Additional file 1: Figure S1) in principle is related to bimodal type of approach in that both detect major separation between outlier and non-outlier groups. This is manifested in the degree of overlap significance between outlier genes identified where Gap shows a stronger agreement with PACK over OS (Fig. 1a). However, unlike bimodal which looks for two normal distributions with distinct means, GAP does not require the outlier group to be Gaussian. Given that the outlier group is often relatively small in size, we believe it is advantageous to circumvent modeling it explicitly in consideration of limited statistical power. The method's utility is highlighted by some well-known cancer genes such as CDK4, APC and EZH2 are only captured by GAP (Additional file 3: Table S2).

The identification of outliers from functional genomic data also helps to uncover potential indications and predictive biomarkers associated with candidate targets (e.g. ovarian cancer and *CCNE1* amplification for *CDK2* inhibitors) that may guide the development of precision medicine strategies. The National Cancer Institute (NCI) has recently launched the Exceptional Responders Initiative to investigate the molecular factors of tumors associated with exceptional treatment responders of cancer patients to drug therapies [42]. Outlier analysis of functional genomics data from large-scale gene silencing provides an opportunity to address similar questions for thousands of genes in parallel using pre-clinical models.

The set of genes with outlier dependency patterns may also provide a useful framework for a global view of tumor cell dependency. Hematological lines have a unique vulnerability pattern that appears more homogenous than their solid tumor counterparts. This may reflect their evolutionary history where leukemias and lymphomas likely require fewer rounds of clonal expansion [43–45] as their precursor cells are already mobile and invasive [46]. It is also possible that some technical factors (e.g. conditions in cell culturing) may contribute to the observed difference. With experimental validation, the grouping of tumor cells by functional dependency could lead to important insights on the design of more sophisticated molecular biomarker strategies of both positive and negative selections for basket trials.

Our apparently counter-intuitive observation that the knockdown of several tumor suppressor genes resulted in striking vulnerability in a subset of tumor cells suggested they could be oncogenic in specific circumstances. This is analogous to the recent finding in the ARID1 family whereby inactivation of *ARID1A* creates a special dependence on the related tumor suppressor gene *ARID1B* [5] and the "abnormal" (*ARID1A*-less) SWI/SNF complex is pro-oncogenic. The presence of additional tumor suppressors like *ARID1B* and *APC* among outlier genes suggests that these are unlikely to be isolated cases and more may have context-dependent dual properties, presenting a challenge to the simple binary classification of cancer genes as either oncogenes or tumor suppressors, and bringing up the possibility of expanding the druggable genome to include tumor suppressors when coupled with an appropriate precision medicine strategy.

While outlier vulnerability analysis of Achilles RNAi sensitivity pattern has successfully uncovered many well-known cancer genes along with their tumor type and genetic biomarker relationships, it also has a few notable misses. *EGFR* and *ERBB2*, two established oncology therapeutic targets, do not appear to harbor a significant exceptional responder group distinct from the population (Additional file 8: Figure S4). We should note that the outlier approach, as any analytical strategy, is contingent upon the quality of the input: technological limitations such as hairpin off-target seed effects [47–51] have an inevitable negative impact on analysis results. Even though we used ATARiS gene-level solutions which have on-target signals greatly amplified as input, our analysis is still not completely immune from this complexity and novel findings in particular need to be experimentally confirmed for proper interpretation. We have described here the application of outlier analysis to only a single large functional genomics dataset; however, the identification of many known cancer genes and relationships provides confidence and proof-of-principle for the general approach.

Whereas we have used Achilles shRNA patterns as an example, outlier analysis should be equally relevant to datasets from alternative functional genomic technologies including siRNA and CRISPR [52–54], as well as to other shRNA screen data. Since outliers are rare events by definition, their detection requires a sufficiently large population as with genome-wide gene expression datasets. Furthermore, for most of the outlier techniques to work effectively, the overall input data (after normalization) should follow a symmetric and preferentially normal distribution without the confounding of excessive technical outliers as discussed above. The outlier analysis approach is not only useful for mining of gene-level values but also lower-level data such as those from individual hairpins where the consistency between outlier responders could provide another means to select for on-target effects.

Conclusions

With the rapid evolution of functional genomic technologies, there is an ever growing demand for analytical strategies to maximize discoveries from the large amounts of data being generated. The current analyses typically focus on genes whose knockdown enhances response in predefined molecular contexts and thus are inherently limited in their ability to reveal new disease-relevant biology. Here we tackle this important conceptual problem and demonstrate one solution by introducing a novel strategy to identify tumor vulnerabilities from functional genomic profiles based on patterns of responsiveness alone. It takes advantage of the observation that oncogene addiction or synthetic lethality generally manifests itself in the exquisite sensitivity of a subset of tumors or cell lines, and therefore is built upon the identification of genes with outlier dropout pattern. We thus expand the utility of outlier analysis, whose application in genomics thus far has been restricted to gene expression data, towards the mining of functional genomic profiles.

We show that genes with outlier vulnerability pattern are strongly and specifically enriched with those known to be associated with cancer and relevant biological processes, demonstrating its utility for the identification of therapeutic targets. The characteristics of the outlier lines can further reveal tumor indications and biomarkers of response associated with candidate targets to guide the development of precision medicine strategies. In addition, it provides a useful framework for a global view of tumor cell dependency, which led to the observation of distinct sensitivity and coherence between solid and hematological malignancies. The counter-intuitive finding of several tumor suppressors with outlier sensitivity patterns challenges the simple binary classification of cancer genes as either oncogenes or tumor suppressors, and generalizes the notion that tumor suppressors could play context-dependent oncogenic roles. Therefore, our novel analytical approach described here offers a valuable alternative means to mine fast-growing functional genomic data in an unbiased manner for discoveries that may lead to the next generation of oncology medicines.

Methods

Achilles data

Project Achilles is a systematic effort aimed at identifying and cataloging genetic vulnerabilities across hundreds of genomically characterized cancer cell lines [4]. The project uses genome-wide genetic perturbation reagents (shRNAs) to silence or knock down individual genes and identifies those genes that affect cell survival. The latest version (2.4.3) was downloaded from the project data portal (http://www.broadinstitute.org/achilles). The file Achilles_QC_v2.4.3.rnai.Gs.gct containing ATARiS [22] gene level scores for 216 cell lines that pass quality control ($p \leq 0.05$) [4] was used as input for our outlier analysis below.

Outlier analysis

(1) PACK

Profile Analysis Using Clustering and Kurtosis (PACK) algorithm implemented in the vabayelMix library of R [13] was applied to Achilles data for outlier identification. We filtered its output for bimodal genes of positive kurtosis (i.e. unequal relative mass), with outlier group more vulnerable to drop-out and containing at least 5 cell lines.

(2) Outlier Sum

The Outlier Sum (OS) algorithm [15] was applied to gene-summarized Achilles data for outlier identification. The outlier-sum statistic for each gene *i* was defined to be "one-sided" as the sum of the values that are smaller than the limit q_{25}(i)-IQR(i):

$$\sum x_{ij} \cdot I\left[x_{ij} < q_{25}(i) - IQR(i)\right]$$

where inter-quartile range (IQR) = q_{75}(i)-q_{25}(i), x_{ij} is the drop-out score for gene i in cell line j and I represents the conditional IF test of whether x_{ij} is lower than this limit (1 if true and 0 if false). Statistical significance was estimated through 10,000 full permutation of the data matrix with outlier-sum statistic calculated in the same manner. We focused on genes with at least 5 outlier cell lines and the false discovery rate (FDR) was determined using the Benjamini–Hochberg procedure [55].

(3) Gap Analysis Procedure (GAP)

We devised a gap-based measure for outlier identification (Additional file 1: Figure S1). Let $x_{i,j}$ be the drop-out value for gene *i* in cell line *j*, reorder the values for each gene so that

$$y_{i,1} \le y_{i,2} \le \ldots \le y_{i,m} \le \tilde{x}_i \le y_{i,m+1} \le \ldots \le y_{i,n}$$

where $\tilde{x}_i$ is the median for gene *i* across cell lines. An outlier sum statistics was computed as following based on gaps (G) between adjacent data points:

$$S_i = \begin{cases} \sum_{j=1}^{k} y_{i,j} \;\; if \;\; \max_{1\le j\le m} G\left(y_{i,j}\right) \ge \alpha\left(y_{i,n} - y_{i,1}\right) \\ 0 \qquad if \;\; \max_{1\le j\le m} G\left(y_{i,j}\right) < \alpha\left(y_{i,n} - y_{i,1}\right) \end{cases}$$

where gap $G(y_{i,j}) = y_{i,j+1} - y_{i,j}$ and $k = \max_{1 \le j \le m} j$ such that $G(y_{i,j}) \ge \alpha(y_{i,n} - y_{i,1})$ For α, we used an arbitrary value of 0.05. Statistical significance was estimated through 10,000 full permutations of the data matrix with the gap-based outlier sum statistic calculated in the same manner. We focused on genes with at least 5 outlier cell lines and the false discovery rate (FDR) was determined using the Benjamini-Hochberg procedure [55].

Most likely functional genetic lesions

To avoid cell culture artifacts and technical biases in cell line collection, mutation hotspots were identified in a systematic and unbiased manner from patient-derived somatic mutation profiles. The Cancer Genome Atlas (TCGA) level 3 mutation data (.maf files) for 20 diverse tumor types (BLCA, BRCA, CESC, COAD, GBM, HNSC, KIRC, KIRP, LAML, LGG, LUAD, LUSC, OV, PAAD, PRAD, READ, SKCM, STAD, THCA, UCEC) were downloaded through firehose (http://gdac.broadinstitute.org/). We filtered for non-silent coding mutations (De_novo_Start_InFrame, De_novo_Start_OutOfFrame, Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins, Missense, Nonstop, Read-through, Translation_Start_Site) and excluded those mapped to pseudogenes. Nonrandom Mutation Clustering (NMC) algorithm [56] was applied to the resulting "Pan-Cancer" mutation profile, with multiple hypothesis testing corrected by the Benjamini-Hochberg procedure [55]. We focused on significant mutation hotspots (FDR ≤ 0.05) of no more than 50 nucleotides. Such length was selected as a compromise for both activating and inactivating missense mutations as activating ones tend to be rather focal [43] while inactivating ones often span a relatively relaxed region/domain [57].

The non-silent coding mutation profiles for cell lines were obtained using Cell Index database (CELLX) [58] based on genetic data compiled from the Cancer Cell Line Encyclopedia (CCLE) [32], Sanger Catalog of Somatic Mutations in Cancer (COSMIC) [59] and Sanger Wellcome Trust Genomics of Drug Sensitivity in Cancer (GDSC) [60]. To select for most likely functional mutations, we filtered for those located at patient-derived hotspots as identified above, as well as loss-of-function (LOF) ones by mutation type of nonsense and frameshift. Copy number segments for cell lines were also obtained using CELLX. Amplifications and deletions for a gene were defined as copy number segments overlapping the gene of at least four copies and at most one copy respectively, if any.

Gene pairs related by families and protein complexes

1012 human gene families together with their associated members were downloaded from genenames.org (http://www.genenames.org/), a curated online repository of HGNC-approved gene nomenclature, gene families and associated resources. Protein complex relationships were obtained from the Comprehensive Resource of Mammalian protein complexes (CORUM) database which provides a resource of manually annotated protein complexes from mammalian organisms including function, localization, subunit composition and literature reference [61, 62]. 1846 such complexes are from human sources. For each pair of outlier gene and its predictive genetic biomarker (see *Enrichment analysis* below), we consider them functionally related if they belong to the same human gene family or protein complex based on the annotations described here.

Clustering analysis

Unsupervised hierarchical clustering was performed with correlation as similarity metric and average linkage as clustering method using Cluster 3.0 software (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm). Non-outliers represent non-sensitive hits and are thus "flattened" to zero. To focus on most variable features, 165 out of the 169 genes with significant outlier sensitivity pattern (Additional file 3: Table S2) were used in the clustering analysis as they each have less than a quarter of total cell lines classified as outliers. The clustering results were visualized using TreeView program [63] and heatmap.2 function of 'gplots' package in R (http://www.inside-r.org/packages/cran/gplots/docs/heatmap.2).

Enrichment analysis

The statistical significance for enrichment of outliers in a tumor (sub)type or with genetic biomarker was calculated as follows:

$$\mathrm{P} = 1 - \sum_{i=0}^{x-1} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}},$$

where M = total of cell lines, K = number of outliers (the union from the three methods), N = number of cell lines from the tumor (sub)type or with likely functional genetic lesion, and x = number of outliers in the tumor (sub)type or with genetic biomarker. The probability of obtaining at least the observed number of common outlier genes from a pair of methods by chance was determined in an analogous manner, where M = total number (5299) of genes with ATARiS consensus solution, K and N = number of outlier genes identified by each method respectively, and x = number of overlapping outlier genes identified by both methods.

Gene signature analysis

Gene signature enrichment analysis was performed by comparing outlier genes with those from the CGC catalogue and MSK-IMPACT™ panel as well as those from the Broad MSigDB's canonical pathway (CP) library. Statistical significance was determined using cumulative hypergeometric probability distribution as previously described [64], where the total number of genes was based on those with ATARiS gene consensus solution [4, 22] and multiple hypothesis testing was corrected by Benjamini-Hochberg procedure [55].

Additional files

Additional file 1: Figure S1. A schematic diagram of the GAP approach. Its matlab implementation is included as Additional file 2. (DOCX 189 kb)

Additional file 2: An implementation of the GAP approach in matlab. (PDF 43 kb)

Additional file 3: Table S1. Tumor type and common subtype classification of Achilles cell lines. (XLSX 16 kb)

Additional file 4: Table S2. The complete list of genes with outlier patterns along with tumor type indication and related predictive biomarker results (sorted by decreasing statistical significance of association), if any. The table indicates for each gene which of the three analysis methods determined it to be associated with an outlier responder pattern. (XLSX 16 kb)

Additional file 5: Figure S2. Unsupervised hierarchical clustering of tumor cells by functional dependency. (A) Grouping of drop-out patterns using outlier genes, where rows correspond to genes and columns correspond to cell lines. Non-outlier genes are colored in grey as they represent non-sensitive hits. The right inset provides a zoom-in view of the yellow cluster with the vast majority of hematological lines. The full size heatmap, dendrograms and labels have been included as Additional file 6. (B) Grouping using outlier genes from the Wnt pathway. (DOCX 1164 kb)

Additional file 6: The complete clustering diagram shown in Additional file 5: Figure S2A including heatmap, dendrograms and row/column labels in full page size. (JPG 10376 kb)

Additional file 7: Figure S3. The outlier pattern for APC. (A) The kernel density plot of ATARiS gene level score for APC and details on the five outlier cell lines most vulnerable to its knockdown. (B) Unsupervised hierarchical clustering of tumor cells by functional dependency on outlier genes from the Wnt pathway, with APC highlighted in yellow. (DOCX 395 kb)

Additional file 8: Figure S4. The ATARiS gene level score distribution for (A) EGFR and (B) ERBB2. A probability density estimate is computed by Gaussian kernel smoothing. (DOCX 49 kb)

Abbreviations

ATARIS, analytic technique for assessment of RNAi by similarity; CCLE, cancer cell line Encyclopedia; CGC, cancer gene census; CORUM, comprehensive resource of mammalian protein complexes; COSMIC, catalogue of somatic mutations in cancer; CP, canonical pathway; CRISPR, clustered regularly interspaced short palindromic repeats; FDR, false discovery rate; GAP, gap analysis procedure; GDSC, genomics of drug sensitivity in cancer; ICGC, International Cancer Genome Consortium; IMPACT, integrated mutation profiling of actionable cancer targets; LOF, loss of function; MSigDB, molecular signatures database; NCI, National Cancer Institute; NMC, nonrandom mutation clustering; OS, outlier sum; PACK, profile analysis using clustering and kurtosis; RNAi, RNA interference; shRNA, short hairpin RNA; siRNA, small interfering RNA; TCGA, the cancer genome atlas

Acknowledgements

We are grateful to Jarek Kostrowicki, Darwin Lee, Kim Arndt, Valeria Fantin, Sapna Vijayakumar, Keith Ching, Shibing Deng, Yu Jean Wang, Pete Olson and Todd VanArsdale for sharing their expertise and insights.

Authors' contributions

PZ, NI and ZZ conceived the study; ZZ, PZ, NI and PR developed computational strategies; ZZ and PZ analyzed the data; PZ and ZZ mapped cell lines to tumor (sub)types; ZZ, PZ and PR drafted and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

All authors are or were employees of *Pfizer Inc.*

Received: 6 November 2015 Accepted: 27 May 2016
Published online: 13 June 2016

References

1. Agrawal N, Dasaradhi PV, Mohmmed A, Malhotra P, Bhatnagar RK, Mukherjee SK. RNA interference: biology, mechanism, and applications. Microbiol Mol Biol Rev. 2003;67:657–85.
2. Boettcher M, McManus MT. Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR. Mol Cell. 2015;58:575–85.
3. Mohr SE, Smith JA, Shamu CE, Neumuller RA, Perrimon N. RNAi screening comes of age: improved techniques and complementary approaches. Nat Rev Mol Cell Biol. 2014;15:591–600.

4. Cowley G, Weir B, Vazquez F, Tamayo P, Scott J, Rusin S, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Nature Scientific Data. 2014;1.
5. Helming KC, Wang X, Wilson BG, Vazquez F, Haswell JR, Manchester HE, Kim Y, Kryukov GV, Ghandi M, Aguirre AJ, et al. ARID1B is a specific vulnerability in ARID1A-mutant cancers. Nat Med. 2014;20:251–4.
6. Wilson BG, Helming KC, Wang X, Kim Y, Vazquez F, Jagani Z, Hahn WC, Roberts CW. Residual complexes containing SMARCA2 (BRM) underlie the oncogenic drive of SMARCA4 (BRG1) mutation. Mol Cell Biol. 2014;34:1136–44.
7. Iyer G, Hanrahan AJ, Milowsky MI, Al-Ahmadie H, Scott SN, Janakiraman M, Pirun M, Sander C, Socci ND, Ostrovnaya I, et al. Genome sequencing identifies a basis for everolimus sensitivity. Science. 2012;338:221.
8. Karrila S, Lee JH, Tucker-Kellogg G. A comparison of methods for data-driven cancer outlier discovery, and an application scheme to semisupervised predictive biomarker discovery. Cancer Inform. 2011;10:109–20.
9. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005;310: 644–8.
10. MacDonald JW, Ghosh D. COPA–cancer outlier profile analysis. Bioinformatics. 2006;22:2950–1.
11. Hellwig B, Hengstler JG, Schmidt M, Gehrmann MC, Schormann W, Rahnenfuhrer J. Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. BMC Bioinformatics. 2010;11:276.
12. Wang J, Wen S, Symmans WF, Pusztai L, Coombes KR. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. Cancer Inform. 2009;7:199–216.
13. Teschendorff AE, Naderi A, Barbosa-Morais NL, Caldas C. PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer. Bioinformatics. 2006;22:2269–75.
14. Bessarabova M, Kirillov E, Shi W, Bugrim A, Nikolsky Y, Nikolskaya T. Bimodal gene expression patterns in breast cancer. BMC Genomics. 2010;11 Suppl 1:S8.
15. Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. Biostatistics. 2007;8:2–8.
16. Wu B. Cancer outlier differential gene expression detection. Biostatistics. 2007;8:566–75.
17. Luo J. Truncated Outlier Robust T-Statistic for Outlier Detection. Open J Stat. 2012;2:120–3.
18. Ghosh D, Chinnaiyan AM. Genomic outlier profile analysis: mixture models, null hypotheses, and nonparametric estimation. Biostatistics. 2009;10:60–9.
19. Mpindi JP, Sara H, Haapa-Paananen S, Kilpinen S, Pisto T, Bucher E, Ojala K, Iljin K, Vainio P, Bjorkman M, et al. GTI: a novel algorithm for identifying outlier gene expression profiles from integrated microarray datasets. PLoS One. 2011;6:e17259.
20. Obikee AC, Ebuh GU, Obiora-Ilouno HO. Comparison of Outlier Techniques Based on Simulated Data. Open J Stat. 2014;4:536–61.
21. Dixon WJ. Analysis of extreme values. Ann Math Stat. 1950;21:488–506.
22. Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, Schumacher SE, Zack TI, Beroukhim R, Garraway LA, et al. ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Res. 2013;23:665–78.
23. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.
24. Asghar U, Witkiewicz AK, Turner NC, Knudsen ES. The history and future of targeting cyclin-dependent kinases in cancer therapy. Nat Rev Drug Discov. 2015;14:130–46.
25. Gutierrez JA, Pan YX, Koroniak L, Hiratake J, Kilberg MS, Richards NG. An inhibitor of human asparagine synthetase suppresses proliferation of an L-asparaginase-resistant leukemia cell line. Chem Biol. 2006;13:1339–47.
26. Li BS, Gu LJ, Luo CY, Li WS, Jiang LM, Shen SH, Jiang H, Shen SH, Zhang B, Chen J, et al. The downregulation of asparagine synthetase expression can increase the sensitivity of cells resistant to l-asparaginase. Leukemia. 2006;20:2199–201.
27. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006;7:96.
28. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6:pl1.
29. Nakayama N, Nakayama K, Shamima Y, Ishikawa M, Katagiri A, Iida K, Miyazaki K. Gene amplification CCNE1 is related to poor survival and potential therapeutic target in ovarian cancer. Cancer. 2010;116:2621–34.
30. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C, et al. Signatures of mutation and selection in the cancer genome. Nature. 2010;463:893–8.
31. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43:D805–811.
32. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.
33. Dupuy AJ, Rogers LM, Kim J, Nannapaneni K, Starr TK, Liu P, Largaespada DA, Scheetz TE, Jenkins NA, Copeland NG. A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. Cancer Res. 2009;69:8150–6.
34. Sauvageau M, Miller M, Lemieux S, Lessard J, Hebert J, Sauvageau G. Quantitative expression profiling guided by common retroviral insertion sites reveals novel and cell type specific cancer genes in leukemia. Blood. 2008;111:790–9.
35. Uren AG, Kool J, Matentzoglu K, de Ridder J, Mattison J, van Uitert M, Lagcher W, Sie D, Tanger E, Cox T, et al. Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. Cell. 2008;133:727–41.
36. Aoki K, Taketo MM. Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. J Cell Sci. 2007;120:3327–35.
37. Chen M, Pratt CP, Zeeman ME, Schultz N, Taylor BS, O'Neill A, Castillo-Martin M, Nowak DG, Naguib A, Grace DM, et al. Identification of PHLPP1 as a tumor suppressor reveals the role of feedback activation in PTEN-mutant prostate cancer progression. Cancer Cell. 2011;20:173–86.
38. Su H, Liu Y, Zhao X. Split End Family RNA Binding Proteins: Novel Tumor Suppressors Coupling Transcriptional Regulation with RNA Processing. Cancer Transl Med. 2015;1:21–5.
39. Zaltsman Y, Shachnai L, Yivgi-Ohana N, Schwarz M, Maryanovich M, Houtkooper RH, Vaz FM, De Leonardis F, Fiermonte G, Palmieri F, et al. MTCH2/MIMP is a major facilitator of tBID recruitment to mitochondria. Nat Cell Biol. 2010;12:553–62.
40. Vanharanta S, Marney CB, Shu W, Valiente M, Zou Y, Mele A, et al. Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer. Elife. 2014;3.
41. Tighe A, Johnson VL, Taylor SS. Truncating APC mutations have dominant effects on proliferation, spindle checkpoint control, survival and chromosome stability. J Cell Sci. 2004;117:6339–53.
42. Mehra N, Lorente D, de Bono JS. What have we learned from exceptional tumour responses?: review and perspectives. Curr Opin Oncol. 2015;27:267–75.
43. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. Science. 2013;339:1546–58.
44. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194:23–8.
45. Laurenti E, Dick JE. Molecular and functional characterization of early human hematopoiesis. Ann N Y Acad Sci. 2012;1266:68–71.
46. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004;10:789–99.
47. Ma Y, Creanga A, Lum L, Beachy PA. Prevalence of off-target effects in Drosophila RNA interference screens. Nature. 2006;443:359–63.
48. Mohr S, Bakal C, Perrimon N. Genomic screening with RNAi: results and challenges. Annu Rev Biochem. 2010;79:37–64.
49. Buehler E, Khan AA, Marine S, Rajaram M, Bahl A, Burchard J, Ferrer M. siRNA off-target effects in genome-wide screens identify signaling pathway members. Sci Rep. 2012;2:428.
50. Sigoillot FD, Lyman S, Huckins JF, Adamson B, Chung E, Quattrochi B, King RW. A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. Nat Methods. 2012;9:363–6.
51. Schmich F, Szczurek E, Kreibich S, Dilling S, Andritschke D, Casanova A, Low SH, Eicher S, Muntwiler S, Emmenlauer M, et al. gespeR: a statistical model for deconvoluting off-target-confounded RNA interference screens. Genome Biol. 2015;16:220.
52. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013;152:1173–83.
53. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet. 2015;16:299–311.

54. Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, Wei W. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature. 2014;509:487–91.
55. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57:289–300.
56. Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng CH. Statistical method on nonrandom clustering with application to somatic mutations in cancer. BMC Bioinformatics. 2010;11:11.
57. Yang F, Petsalaki E, Rolland T, Hill DE, Vidal M, Roth FP. Protein domain-level landscape of cancer-type-specific somatic mutations. PLoS Comput Biol. 2015;11:e1004147.
58. Ching KA, Wang K, Kan Z, Fernandez J, Zhong W, Kostrowicki J, et al. Cell Index Database (CELLX): a web tool for cancer precision medicine. Pac Symp Biocomput. 2015;10–19.
59. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011;39:D945–950.
60. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41:D955–961.
61. Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegele B, Schmidt T, Doudieu ON, Stumpflen V, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 2008;36:D646–650.
62. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. 2010;38:D497–501.
63. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95:14863–8.
64. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22:281–5.