Supplement for PLATYPUS: A Multiple–View Learning Predictive Framework for cancer drug sensitivity prediction

S1. Alternative Multiple-View Learning Approaches

Alternative forms of MVL have been applied to other bioinformatics problems. For example, multiple kernel learning (a form of multiview learning) has been used in bioinformatics problems in prostate chemical recurrence prediction,[7] and was used in several top-performing methods in a DREAM competition to predict drug sensitivity.[4] While previous work has used multiple kernel learning to combine different biological data platforms,[9,10] this approach is a type of single view classifier because it only uses one form of prior (*e.g.* pathways). Another approach in the same challenge uses a form of multiview learning that ignores samples with missing data, using canonical correlation analysis (CCA) to combine data platforms. While using CCA is appropriate in situations where data are highly correlated,[11] in biological problems the data is noisy and different perspectives of the data are dissimilar, leading to low–correlated views and poor classifier performance. Furthermore, neither multiple kernel nor CCA–based multiview methods gracefully handle missing data.

Each of these MVL frameworks have a unique approach to data integration. CCA is an example of early integration because it combines the data before supervised analysis. Multiple kernel methods integrate during model training and operate as intermediate integration stages. Ensembles and stacked learning approaches, such as PLATYPUS, use late–stage integration, where individual models are trained on separate data then integrated after training. This approach is more flexible but requires a second training phase. While each approach has it's benefits, we chose to use late stage integration for the increased flexibility in handling missing data.

In the situation where multiple data types are available on a common set of samples, one may be tempted to combine the dataset into one set of features and then apply a feature selection strategy to the data (e.g. Random Forests). However, a combined dataset containing all features would introduce many missing values. Separating the features into views minimizes the missing data problem, allowing them to be handled within each data type. Using one large combined feature set would force a developer to either restrict to a small set of samples for which all data is available or would require a sophisticated method for handling large amounts of missing data. PLATYPUS currently handles the missing data in the integration step of the different classifiers and thus avoids introducing additional missing data for samples measured only on a subset of platforms.

S2. Data in detail

CCLE data was downloaded from *www.broadinstitute.org/ccle/home.*

Mutation: CCLE includes mutation data in the form of Single Nucleotide Polymorphism (SNP) and insertion or deletion (Indel) events for 1,651 genes. These were assessed by targeted massively parallel sequencing and later filtered, *e.g.* for presumably neutral variants or common polymorphisms. Additionally, 392 mutations in 33 genes known to be associated with cancer[27] were assessed by mass spectrometric genotyping. SNPs and Indels were combined into a set of non-silent mutations that include all events changing the amino acid composition of the resulting protein, including Indels or missense SNPs in the coding region, splice site, and stop or start codon alterations.

Expression: Gene expression data measures expression over 18,900 genes using Affymetrix U133 plus 2.0 arrays, converted to single gene values by Robust Multi-array Average (RMA) and quantile normalization.

CNV: Copy Number Variation (CNV) data covers 23,316 genes and was determined using Affymetrix SNP6.0 arrays, and normalizing the values with the most similar HapMap normal samples. In the CNV data, some genes have the same value for each cell line because they lay on the same genome segment varying in copy number. Each of these gene sets was merged into one feature in order to reduce redundancy in the data, resulting in 20,247 features.

Clinical: Sample annotation data for the CCLE cell lines contains the gender of the cancer patient and information about the cancer origin (*i.e.* 24 different tissue types, 21 histology types, and 67 histology subtypes).

Drug Sensitivity: There is drug response data to 24 anti-cancer drugs for about 50% of the cell lines in CCLE. A fitted dose-response curve from eight measurements is given, together with the inferred values for EC50, IC50, and Activity Area (ActArea), the area over the dose-response curve (see Fig. 2b from Barretina *et. al.*[2] for definition). ActArea was used for all analyses in this work for three reasons:

(1) ActArea captures more information about the dose-response curve than a single point like IC50 or EC50, *i.e.* the angle of the curve and initial points of sensitivity changes.
(2) ActArea is always given. EC50 in contrast is set to NA if no sufficient response was measured with the maximal tested dose.
(3) ActArea has no artificial values, whereas IC50 is set to the maximal tested dose if no response was measured.

S3. Views

S3.1. *Baseline Views*

Two major types of views are used in PLATYPUS – baseline and interpreted. Baseline views are data platforms that do not incorporate prior knowledge. In this study, we use four baseline views: copy number, non–silent mutation, RNA expression, and clinical/phenotype information. The mutation view is composed of gene–level binary features representing the presence of a non–silent mutation in a specific gene in a particular sample. Copy number and expression data are normalized continuous gene–level values (Section S2). Data in the RNA expression and copy number views is reduced to the 5,000 genes with the highest variance over all cell lines.

The Sample- and Patient-Summary (SPS) view, is a collection of features based on clinical/phenotypic data such as tissue, histology, and gender. For the CCLE cell lines, this information includes, for example, the gender of the cancer patient and information about the cancer origin. In total, 24 different tissue types, 21 histology types, and 67 histology subtypes) are represented. Additionally SPS includes two features representing overall genomic instability of the cell line: the total number of mutated genes and the sum over all absolute CNV values, in each cell line. The values are quantized into 'low', 'medium', and 'high' instability indexes.

S3.2. *Interpreted Views*

Interpreted views incorporate prior knowledge from scientific literature. Biological experiments of the past century have elucidated the underlying rules of biological processes and a plethora of databases exist that have collected this information. For example genes operate in pathways (multi–protein complexes, signaling cascades, transcriptional regulons, chromatin domains, *etc.*) and gene modules that summarize the activity in groups of genes in different aspects. The use of pathways as gene sets has been shown to be effective at increasing interpretability in the cancer setting, thus motivating their use to create views. Additionally, 'master regulator' analysis that finds transcription factors responsible for observed gene expression changes, which has been shown to identify key transcriptional regulators driving the cancer phenotype, can also serve as an informative view. Using such features could potentially highlight functional changes in the samples which are not necessarily clear at the gene level.

PLATYPUS creates views from gene sets in two major ways: 1) By creating summaries or 2) by restricting the data to the subset of genes in a given set. These two view creation approaches are briefly described in the next sections.

Views Based on Gene Set Summaries. PLATYPUS converts the gene–level data to higher interpreted information by creating views based on given gene sets representing pathways or gene modules. In this study, We tested the use of views derived from gene expression data combined with prior pathway knowledge. The Molecular Signatures Database (MSigDB)[13] was used as the source of pathway–specific gene sets as it contains many pathways relevant to cancer processes. Simple aggregation statistics are used as views in this study to summarize the expression levels of genes belonging to the same pathway including mean, median, and variance. In addition, We included kurtosis to detect when gene sets contain one or multiple genes with extremely different expression levels compared to the other members of the pathway, which could be a sign of a perturbation to the pathway's regulation.

The spatial form of the genome and chromatin structure are closely tied to the cell-of-origin of tumors, which has a major influence on the manifestations of the cancer's progression and response. We created a view based on physical structural proximity of genes in the genome. A recent study found that chromatin interaction domains are both highly stable and have few boundaries that differ between cell types.[5] It also deconvolves tissue–specific noise, which have been strongly correlated with expression.[34] Inclusion of the Drug–Gene Interaction database[8] clarifies the effects of mutations on the response of cell lines to chemical agents. Both of these

Table S1. Median AUC for single view CNV tests

Drug	Median AUC
AEW541	0.502
AZD0530	0.506
AZD6244	0.524
Erlotinib	0.498
L–685458	0.505
Lapatinib	0.528
LBW242	0.532
Nilotinib	0.519
Nutlin–3	0.508
Paclitaxel	0.578
Panobinostat	0.630
PD–0325901	0.526
PD–0332991	0.530
PF2341066	0.503
PHA–665752	0.521
PLX4720	0.505
RAF265	0.544
Sorafenib	0.519
TAE684	0.495
TKI258	0.558
Topotecan	0.597
X17–AAG	0.504
ZD–6474	0.495
Mean	0.527

databases provide insights into the function of drug sensitivity, and knowing which genes a drug interacts with helps focus attention to that subset of genes and their interactors.

Views Based on Restricted Gene Set Subsets PLATYPUS creates views as restricted subsets of the feature data based on provided gene sets. The features in these views remain unchanged from their original. Limiting to a relevant pathway may guide a predictor to identify informative feature combinations that may be otherwise missed due to the high-dimensional nature of the problem. The complete set of gene sets used to create subset views in this paper are described in detail in Section S4. We also created these views with CNV data, however they are excluded due to poor predictive power (average AUC 0.527, Table S1). In this paper, gene set views are constructed from expression data and mutation data (Fig. S3).

S4. Biological Priors

S4.1. *Biological Gene Sets*

Metabolic Enzymes: The metabolic enzymes gene set was created by collecting all genes in the CCLE data belonging to the Cytochrome P450 (CYP) family. CYP proteins are the key players in drug metabolism; They deactivate or facilitate the excretion of most drugs, but they also transform many drugs into their active form.[33] There are 53 CYP proteins in the

CCLE expression data.

Multi-Drug Resistance Proteins: Expression data was subset to a list of multi-drug resistance proteins based on.[32] All 12 defined proteins are present in the data set.

Drug Targets: This view includes all proteins targeted by the 24 anti-cancer compounds in the CCLE data set. The information about drug - protein interactions was collected from DrugBank,[31] a recent review of drug targets,[16] the Drug Gene Interaction Database (DGIdb),[8] and manual literature curation for drugs without an annotated target in the sources named before (Table S2). In total, 142 genes found to be drug targets were present in the gene expression data set. In addition to the expression-based view, this prior was also used to create a view using the mutation data, in which 82 of the drug targets are present.

Table S2. Drug targets curated from literature

CCLE Compound	Target	Source
L-685458	PSEN1	[18]
L-685458	PSEN2	[18]
LBW242	XIAP	[17]
Nutlin-3	MDM2	[19]
PHA-665752	MET	[29]
TAE684	ALK	[20]

Chromatin-Modifying Enzymes: This gene set includes chromatin-modifying proteins.[30,35] It contains 65 proteins, of which 56 are present in the gene expression data.

Druggable Genes: The druggable genes view was created from DrugBank,[31] a recent drug target review,[16] cell surface proteins as defined in,[21] membrane proteins and genes on the druggable genome list from DGIdb,[8] a manually curated list of kinases (Table S2), and the Therapeutic Target Database (TTD,[22]). In contrast to the Drug Targets view, the proteins in this set are not limited to the 24 CCLE compounds. Proteins that are not a target of any existing drug, but have the characteristics to serve as one, are included. A total of 4,632 genes from this gene set are present in the gene expression data.

Essential Genes: The information about essential genes in cancer cell lines was retrieved from Project Achilles,[23] an effort to identify genes having an effect on cell viability by using short hairpin RNA (shRNA) screens. Two versions of Achilles were merged in order to maximize the overlap with the cell lines in CCLE: Achilles v.2.11 and v.2.4.3. The 30 most essential genes for each cell line present in both CCLE and Achilles were retrieved. CCLE expression data was subset to the union of these genes resulting in 2,064 features for this view.

S4.2. *MSigDB Gene Sets*

The Molecular Signatures Database (MSigDB)[15] provides biological gene sets in different collections. Median, variance, and kurtosis values of gene expression in each gene set was calculated and defined as a feature. For using CCLE mutation data with MSigDB gene sets, the enrichment of the mutated genes of a cell line in a gene set was tested using hypergeometric distribution (R function *phyper*). The following MSigDB collections were chosen:

Hallmark Gene Sets: A collection of gene sets created from overlapping gene sets. It features reduced noise and redundancy and contains 50 gene sets.

Motif Gene Sets: 836 gene sets containing genes that share conserved cis-regulatory motif.[24]

Transcription Factor Targets: A gene set contains all genes sharing a transcription factor binding site defined by a TRANSFAC record.[25] There are 615 gene sets in this collection.

Positional Gene Sets: Gene sets corresponding to the position of genes on the human genome regarding chromosome and cytogenetic band. The collection holds 326 gene sets.

Oncogenic Signatures: Signatures of 189 cellular pathways which are often dis-regulated in cancer.

Immunologic Signatures: The 1,910 gene sets represent cell states and perturbations within the immune system.

S4.3. *Drug Target Gene Sets*

For each drug target defined in Section S4.2, all genes occurring in at least one gene set together with the target gene were unified to build one drug target gene set. The MSigDB collections Hallmark, Oncogenic, and Immunologic were used separately. As before, median, variance, and kurtosis of the expression values were calculated for each drug target gene set and used as features.

S4.4. *Regulator Activity by Viper*

Virtual Inference of Protein-activity by Enriched Regulon analysis (Viper) is a tool to transform gene expression features into regulator activity.[26] It takes as input gene expression, a regulon (bipartite regulation network of regulators, e.g. transcription factors), and the genes that are regulated by them. Here, a general regulon called 'multinet'[28] and the CCLE expression data were used. Viper was run in R as part of the Bioconductor project.[26]

S5. Implementation

There are two main inputs to PLATYPUS: (1) the binary outcome labels (*e.g.* 'sensitive' or 'non–sensitive') of the labeled samples and (2) the data view objects. View objects contain the feature matrix, classifier type, optimized parameters for the algorithm (optional), and weight for that view (optional).

In each iteration, views are used to predict labels for unlabeled data. View votes are weighted by either accuracy (as described in Section 2.3) or by a user–provided value. Weights can be static or updated at each iteration, as specified by the user. PLATYPUS automatically handles predictions on samples with missing data. To decide when to stop the process of label learning, the user can define a maximum number of iterations and/or change the necessary vote agreement by providing λ. When making new predictions using the trained PLATYPUS model, users may select their preferred iteration and use that intermediary PLATYPUS model instead of the final iteration's model. Learned feature weights from each view, as well as the models' AUCs, can be extracted for each iteration.

S5.1. *View Creation*

PLATYPUS views were created using the package's view–creation functions. Additionally there is a function for creating view configuration files based on an existing view, which is useful when running PLATYPUS from the command line. Configuration files contain parameters for a view, which vary based on classifier type. At time of publication PLATYPUS supports 3 classifiers: support vector machine (SVM),[6] elastic net (EN),[14] and random forest (RF).[3] Configuration files and views can be created using the function *generate.single.view*, which performs a parameter sweep to find the highest AUC parameters over 100 tests given a data set, model type, and outcome labels.

Fig. S3 shows the single–view tests on the full CCLE data. At this stage a user can decide which models to include based on their performance. Note that it is not always best practice to remove a low–performing view, since view performance may either improve in later iterations or be useful in downstream analysis.

Features must be processed before creating a view configuration file. There are two functions for view feature processing/creation. For baseline views, the user can subset the data to a user–specified number of features. Interpreted views can be created using the *generate.feature.data.summary* function, which combines a biological prior with a feature set. This function takes a dataset, a prior knowledge module network, and a summary metric (*e.g.* median, variance, kurtosis, max), and outputs a sample–by–module feature matrix. To minimize missing data issues, there is an optional parameter for the minimum number of features per module required for that module to be included in the view (by default 3). Once features and view objects are created, they can be used in PLATYPUS.

S5.2. *PLATYPUS Output*

After training, PLATYPUS returns an object containing information from each iteration. This includes label lists for both the labeled and unlabeled samples, AUCs for each view, and the full PLATYPUS model. PLATYPUS objects can be used to make predictions on new samples or to extract features (plus their weights, if relevant) from the views. Label–learning validation results can also be visualized as shown in Fig. S5(b). While unlabeled samples cannot be included in the AUC calculation because their true labels are unknown, AUC based on the labeled samples can be used as a proxy (Fig. S5(a)).

S5.3. *Label learning Validation*

In order to validate the learning process, we introduce *label learning validation* (LLV, Fig. S5(b) and S6). Similar to cross–validation, LLV masks a subset of the labels, then trains the model using the remaining labeled samples. Masked samples are treated as unlabeled data, for which PLATYPUS then tries to infer the labels. LLV compares the learned labels to the (masked) true labels. Views are trained using the $(k-1)$ folds of labeled samples and inferred labels for unlabeled samples from earlier iterations. After performing all k folds, all labeled samples have an additional learned label, except in cases where it could not be learned either due to strong disagreement between the views or extensively missing data.

Fig. S5(b) shows a label learning visualization of PD–0325901. PLATYPUS has correctly relearned the majority of labels. This example would have benefited from stopping at an earlier iteration of PLATYPUS, since the majority of incorrect labels are learned during later iterations. The dashed grey line in parts *a* and *b* show a recommended stopping point, which would be selected by the user based on the LLV output. LLV is a useful tool for choosing an appropriate maximum number of PLATYPUS iterations.

LLV can indicate PLATYPUS' confidence in the predicted labels and how PLATYPUS does not force a label on samples for which the single views disagree substantially. Fig. S6 shows the LLV visualization for each of the 24 CCLE drugs. For each drug, PLATYPUS successfully learns the majority of sample labels correctly. The learning processes differ between drugs and we postulate that there is no globally optimal number of iterations. LLV helps the user see how PLATYPUS performs on the labeled data, which can then be used to extrapolate its performance on unlabeled data.

S6. Mutations Correlate with Drug Sensitivity

For every drug–gene pair in CCLE we used a two–sided t–test to identify markers of drug sensitivity, using Bonferroni multiple hypothesis correction (Table S3). However, a simple t–test approach fails to identify co–occurring mutation interference in the results. PLX4720 is an example: While it does not target KRAS, there is a significant correlation between sensitivity and mutation status. This is explained by BRAF–mutated cell lines, which are sensitive to PLX4720 and which confound the results.

Two different types of BRAF–mutated cell lines interference are seen in this data (Fig. S1): First, KRAS–PLX4720 appear to be significantly associated due to the strong responses of BRAF–mutant cell lines making up a large portion of the KRAS–wildtype group. Because of these cell lines, the KRAS–mutated cell lines appear incorrectly to have a significantly lower response than the KRAS–wildtype cell lines. Similarly, other genes erroneously appear to be significantly associated with PLX4720. Second, KRAS–mutant cell lines incorrectly do not appear to be sensitive to PD–0325901 and AZD6244. As with PLX4720, this is due to the large number of BRAF–mutants in the KRAS–wildtype cell lines.

To validate this theory, we repeated the experiment but excluded all data from cell lines with a BRAF mutation. Excluding BRAF–mutated cell lines leads to KRAS–PD–0325901 and KRAS–AZD6244 correlation and removes the erroneous correlation between KRAS and PLX4720. Thus is it inadvisable to make drug treatment decisions using a single gene mutation status, since it is insufficient to make conclusions about sensitivity to a drug. More sophisticated methods, which take many features into account, are important for identifying co–occurring events which confound sensitivity predictions.

Table S3. Significant Gene–Drug Pairs

gene	**drug**	**p–value**	**annotated**	**alternative name**
BRAF	PD–0325901	$1.08e{-7}$	**yes**	
BRAF	AZD6244	$2.07e{-7}$	**yes**	Selumetinib
PPARGC1A	PD–0332991	$3.07e{-6}$	no	Palbociclib
RB1	PD–0332991	$6.87e{-6}$	no	Palbociclib
CDKL2	PLX4720	$7.88e{-6}$	no	Vemurafenib
BRAF	PLX4720	$1.45e{-5}$	**yes**	Vemurafenib
FES	Nilotinib	$5.03e{-5}$	no	
KRAS	PLX4720	$1.01e{-3}$	no	Vemurafenib
DBF4	PF2341066	$1.88e{-3}$	no	Crizotinib
BAI1	PLX4720	$2.01e{-3}$	no	Vemurafenib
MMP8	PD-0332991	$5.94e{-3}$	no	Palbociclib
RHPN2	Erlotinib	$9.18e{-3}$	no	
FES	TKI258	$1.16e{-2}$	no	
CLTC	PD-0332991	$1.53e{-2}$	no	Dovitinib
ERCC6	PF2341066	$2.95e{-2}$	no	Crizotinib
ALK	PF2341066	$2.96e{-2}$	**yes**	Crizotinib
ROCK2	PLX4720	$4.69e{-2}$	no	Vemurafenib

S7. Identifying Patients with an Aggressive Subtype of Prostate Cancer

Treatment to reduce or block testosterone in men with prostate cancer is often effective but tumors progress in some patients. Two recent projects funded by Stand Up To Cancer – a West Coast Dream Team (WCDT) and an East Coast Dream Team (ECDT) – collected samples of advanced, treatment resistant prostate cancer and characterized their genomes and transcriptomes with DNA and RNA sequencing. Out of 46 total samples with resistant cancer, the WCDT identified 10 with a rare type of histology that they named treatment-emergent small-cell neuroendocrine prostate cancer (t-SCNC).[1] To see if PLATYPUS could generalize to a completely new domain, we asked the model to identify small cell disease in the ECDT cohort of 189 samples for which no histology calls were available, but 117 samples had gene expression data.

Before training PLATYPUS to predict t-SCNC from non-t-SCNC using the WCDT samples, the combined WCDT&ECDT RNA expression data was batch corrected using the ComBat algorithm.[12] From these data, 9 single views were built as described in Table S4. The model for each view was trained 100 times using the same folds for cross-validation. Three views had an average AUC higher than 0.80 and were selected to train the PLATYPUS model; these views were 'Hallmark Gene Sets', 'Expression 5k', and 'Chromatin–Modifying Enzymes'.

PLATYPUS trained for eight iterations with the user–specified learning threshold λ set to 70%. The initial WCDT training set included ten small cell and 36 adenocarcinoma samples. After label learning, all of the training samples were predicted correctly. Whereas, in the initial ensemble model, there was one mislabeled sample (Fig. S9).

Applying the fully trained PLATYPUS model to the ECDT data predicted seven small cell (MO-1012, MO-1118, MO-1215, SC-9001, SC-9031, SC-9066, and SC-9096) and 109 adenocarcinoma samples. One sample, TP-2061, remained unlabeled. Comparing these predictions

Table S4. Single views considered for the combined run of ECDT&WCDT data

View Name	Type	# Features	Origin	AUC
Hallmark Gene Sets	s	50	MSigDB	0.87
Expression 5k	gs	5,000	most varying genes	0.84
Chrom. Mod. Enzymes	gs	65	Allis et al 2007	0.81
Oncogenic Signatures	s	189	MSigDB	0.79
Positional Gene Sets	s	343	MSigDB	0.77
Druggable Genes	gs	4,963	DrugBank,DGIdb,TTD	0.75
Transcr. Factor Targets	s	615	MSigDB	0.72
Motif Gene Sets	s	836	MSigDB	0.70
Immunologic Signatures	s	1,910	MSigDB	0.62

AUC is the average calculated from 100 cross validation runs, each with a unique sets of folds. The same fold sets were used on all views. Views with greater than 0.8 AUC are included in the PLATYPUS experiments. Type labeled as ‘s’ for summary and ‘gs’ for gene set, see Section S4 for a detailed description of the biological priors.

to the now available Neuroendocrine Prostate Cancer (NEPC) classification of the ECDT samples reveals that both of the ECDT samples with neuroendocrine differentiation were indeed classified as small cell by PLATYPUS (Table S5). Furthermore, all of the samples predicted by PLATYPUS to be Adenocarcinoma are also classified by the ECDT as Adenocarcinoma. Finally, four samples listed as Adenocarcinomas have been reclassified as small cell by PLATYPUS and thus may represent an important point of disagreement that further clinical investigation may shed light upon.

Table S5. Comparison of PLATYPUS predictions to histology annotation in the ECDT data set.

PLATYPUS prediction	total	NEPC Class A	NEPC Class B	unclassified
Adenocarcinoma	109	98	0	11
Small Cell	7	4	2	1
No prediction	1	1	0	0
	117	103	2	12

NEPC Class A = usual high grade prostatic adenocarcinoma; NEPC Class B = high grade prostatic adenocarcinoma with neuroendocrine differentiation.

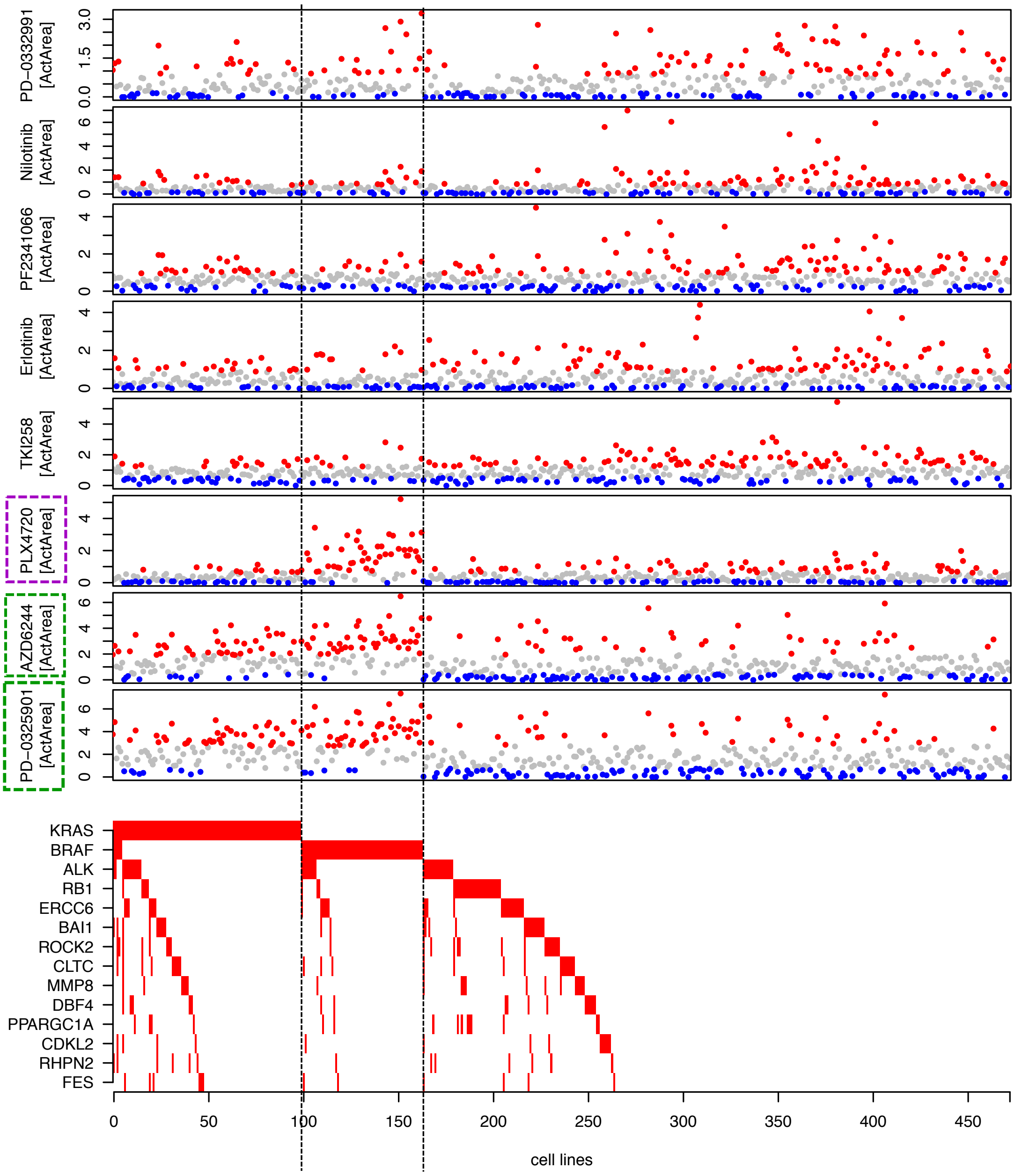

Fig. S1. Mutation status and drug response of each cell line for significant (Bonferroni corrected) compound-gene pairs; with a focus on KRAS mutation state and compound-KRAS combinations. Heatmap on the bottom shows the mutation status of different genes (y-axis) in the cell lines (x-axis), where red represents a non-silent mutation. The cell lines are ordered according to the mutation state in the genes along the y-axis, starting from the top. The vertical lines separate cell lines with a mutation in the first two genes (KRAS and BRAF, respectively) from cell lines with the wildtype genes. Dot-plots on top show the according drug response of the cell lines in different compounds. Drug response coloring reflects binary response (red=sensitive (top quartile), blue=insensitive (bottom quartile), grey=intermediate (second and third quartile)). For the t-tests the actual ActArea value was used. PLX4720 has a significant p-value in combination with KRAS is circled in (purple), but KRAS is not an annotated target of PLX4720. When removing the BRAF-mutant cell lines from the data set, PD-0325901 and AZD6244 are significant towards KRAS mutation (green).

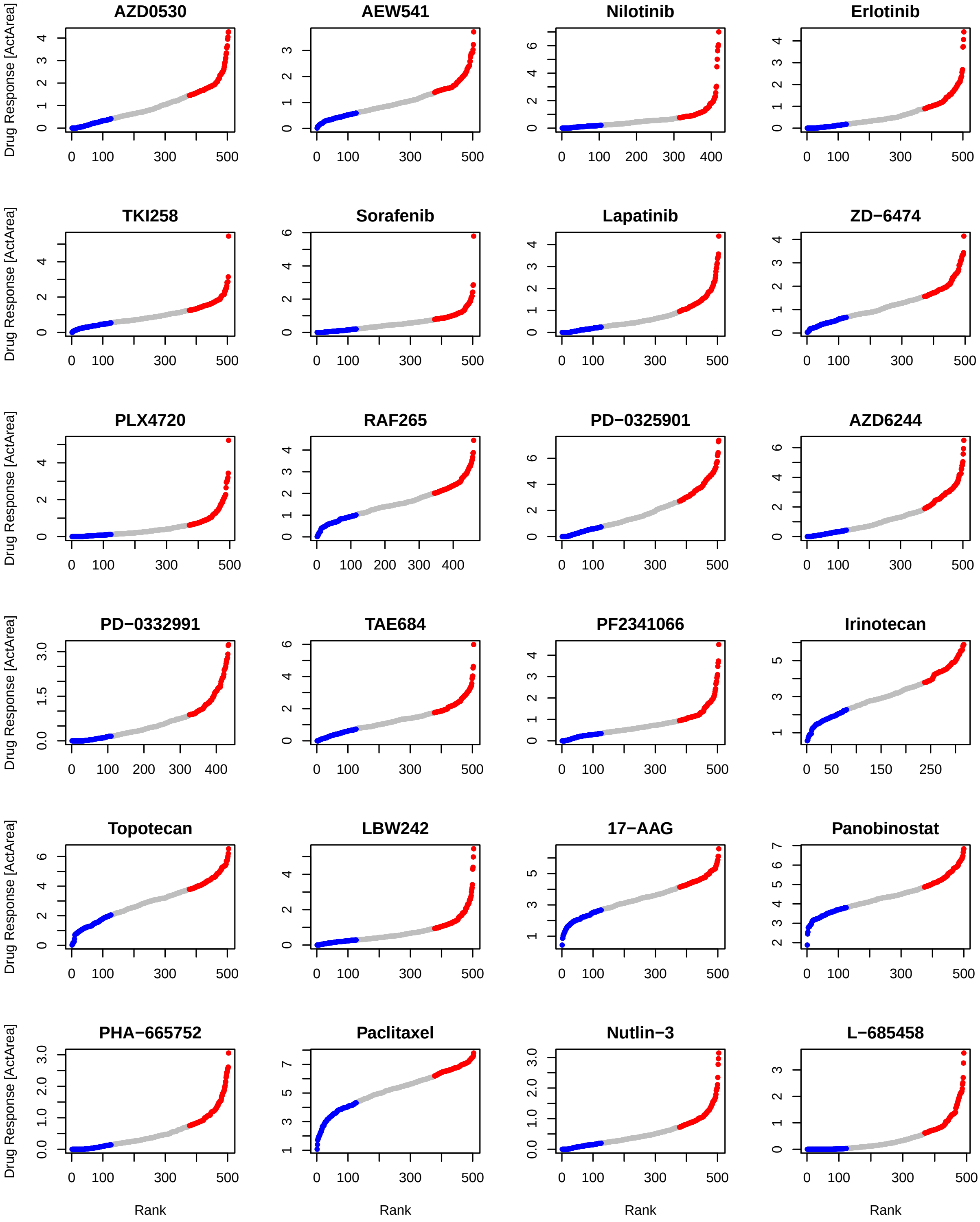

Fig. S2. The ranked ActArea values of each CCLE cell line for the 24 CCLE compounds. Blue dots are cell lines labeled as ‘non–sensitive’ for the correspondent drug, red ones are labeled ‘sensitive’, gray ones ‘intermediate’. The number of cell lines in the non–sensitive class corresponds to the bottom 25% of cell lines the drug response was measured for, the ‘sensitive’ class to the top 25%.

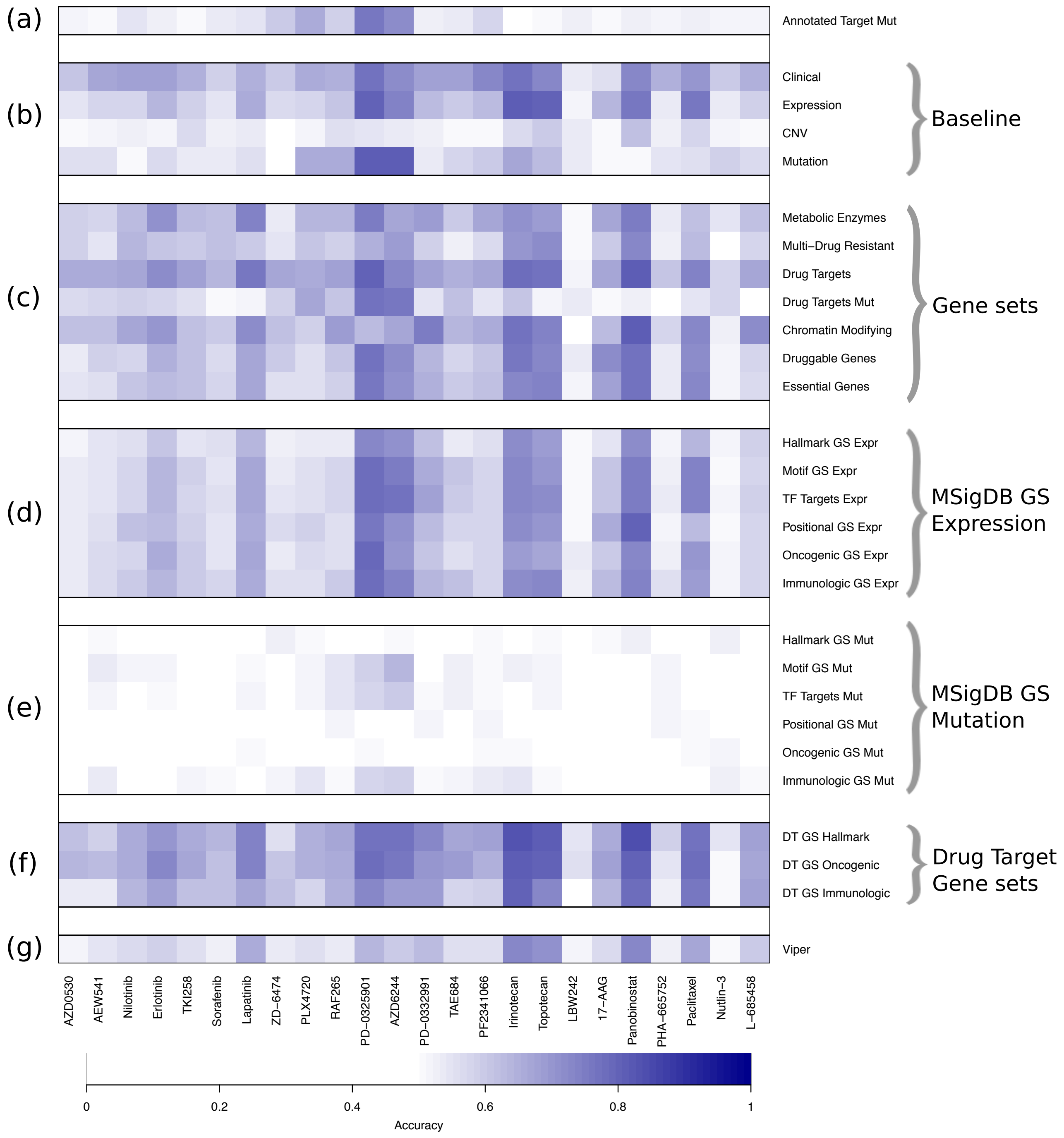

Fig. S3. AUC for each view when predicting sensitivity to each drug in CCLE. Grouped by data type. The cross-validated AUC of single views with their optimized parameter settings. All values ≤ 0.5 (AUC of a random predictor) are shown in white. The simple Annotated Target Mutation predictor (Section S6) is shown in A. The following single views are grouped according to Sections S4 and S2. GS = Gene Set; DT = Drug Target; Expr = Expression; Mut = Mutation.

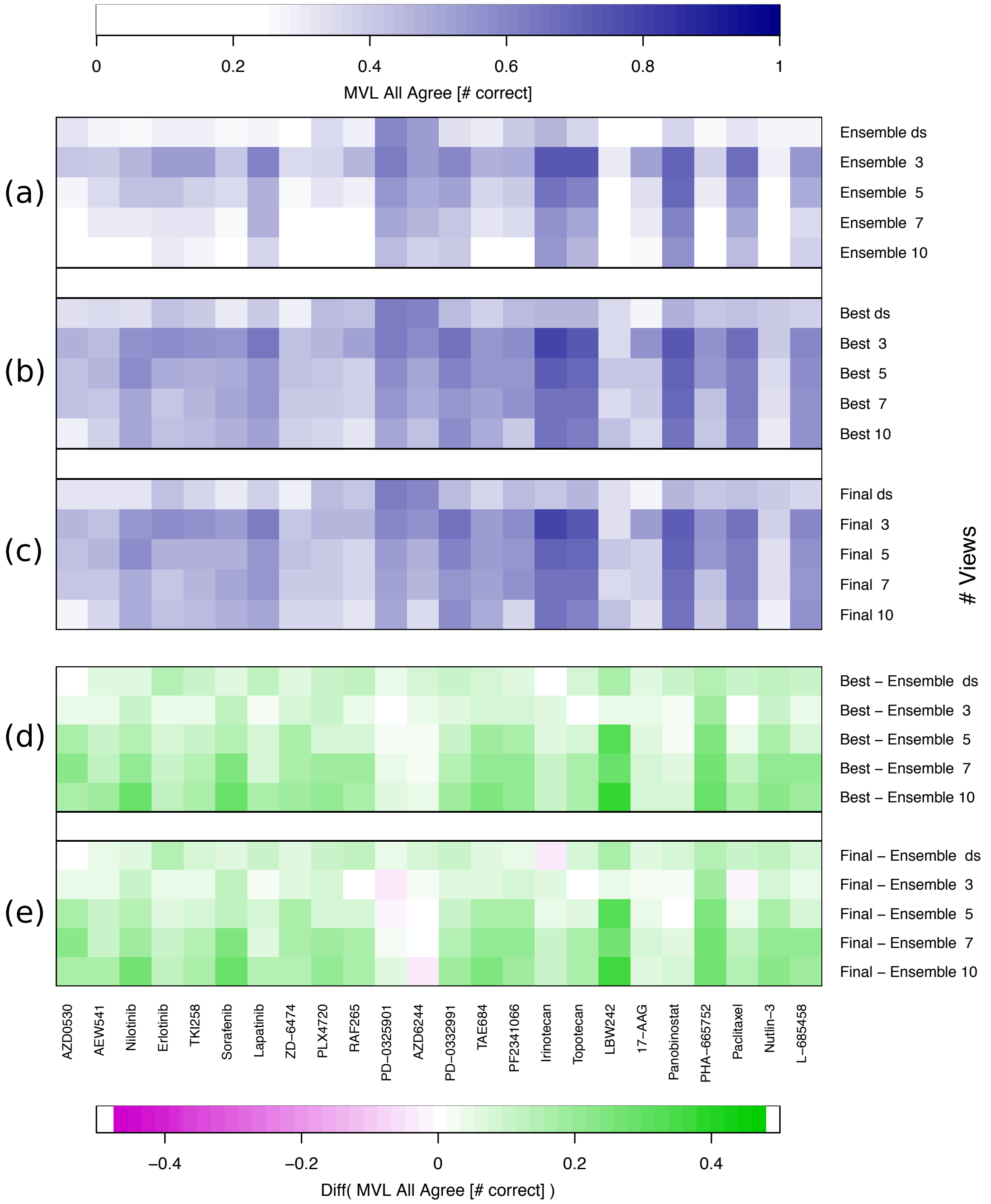

Fig. S4. Cross-validated number of correctly predicted samples (balanced accuracy × coverage) of PLATYPUS for 100% agreement predictions. (a) ‘Ensemble’ represents the first iteration of the MVL algorithm, in which no inferred labels have been added. (b) ’Best’ is PLATYPUS iteration with the highest AUC. (c) ‘Last’ is the model from the final PLATYPUS iteration. Inferred labels were added until 75% agreement. (d-e) Comparison between the different MVL models by subtracting the Ensemble performance from the (d) Best and (e) Final performances. Each compound was predicted with the data-specific (ds) views (SPS, Mutation, Expression), and with the 3, 5, 7, and 10 most accurate interpreted single views.

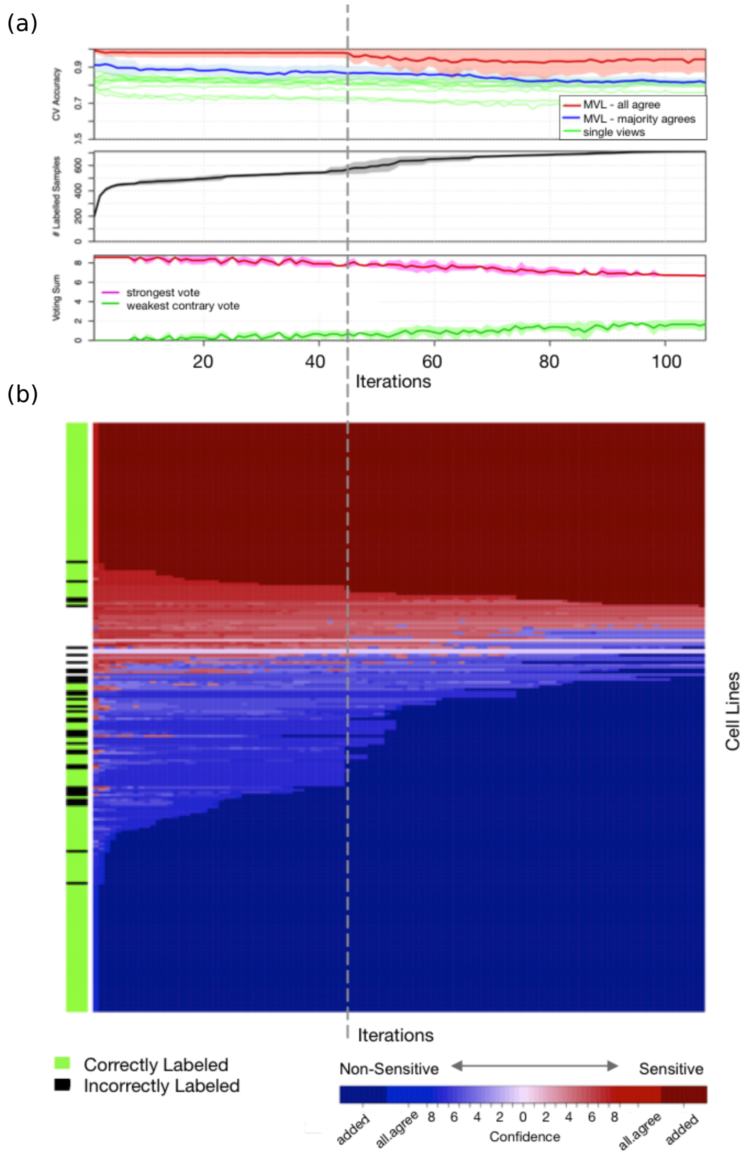

Fig. S5. (a-b) Dashed vertical line shows user–defined stopping point for the method, where the overall disagreement in predictions between the views has started to increase and before the model area under the receiver-operator curve (AUC) starts to significantly decrease. (a) Cross-validation accuracy mean (solid line) and standard deviation (colored area) plotted at each iteration (x-axis). Top plot: the prediction accuracy for single views (green lines), PLATYPUS ensembles with majority (75%) agreement (blue line), or all view votes agreeing (red line) on unlabeled samples. Middle plot: For each iteration, the number of samples for which labels have been learned (y-axis). Bottom plot: The votes summed up for the unlabeled sample with the highest vote (pink line) and smallest vote (green line) at each iteration across the cross-validation folds. (b) Label learning progress over successive iterations (x-axis) showing confidence of predictions for each unlabeled cell line (y-axis). Success or failure of the method to assign the correct label to each cell line shown in first column (green, correct; black, incorrect). Darker color indicates higher vote confidence across the views for either the non-sensitive (blue) or sensitive (red) class.

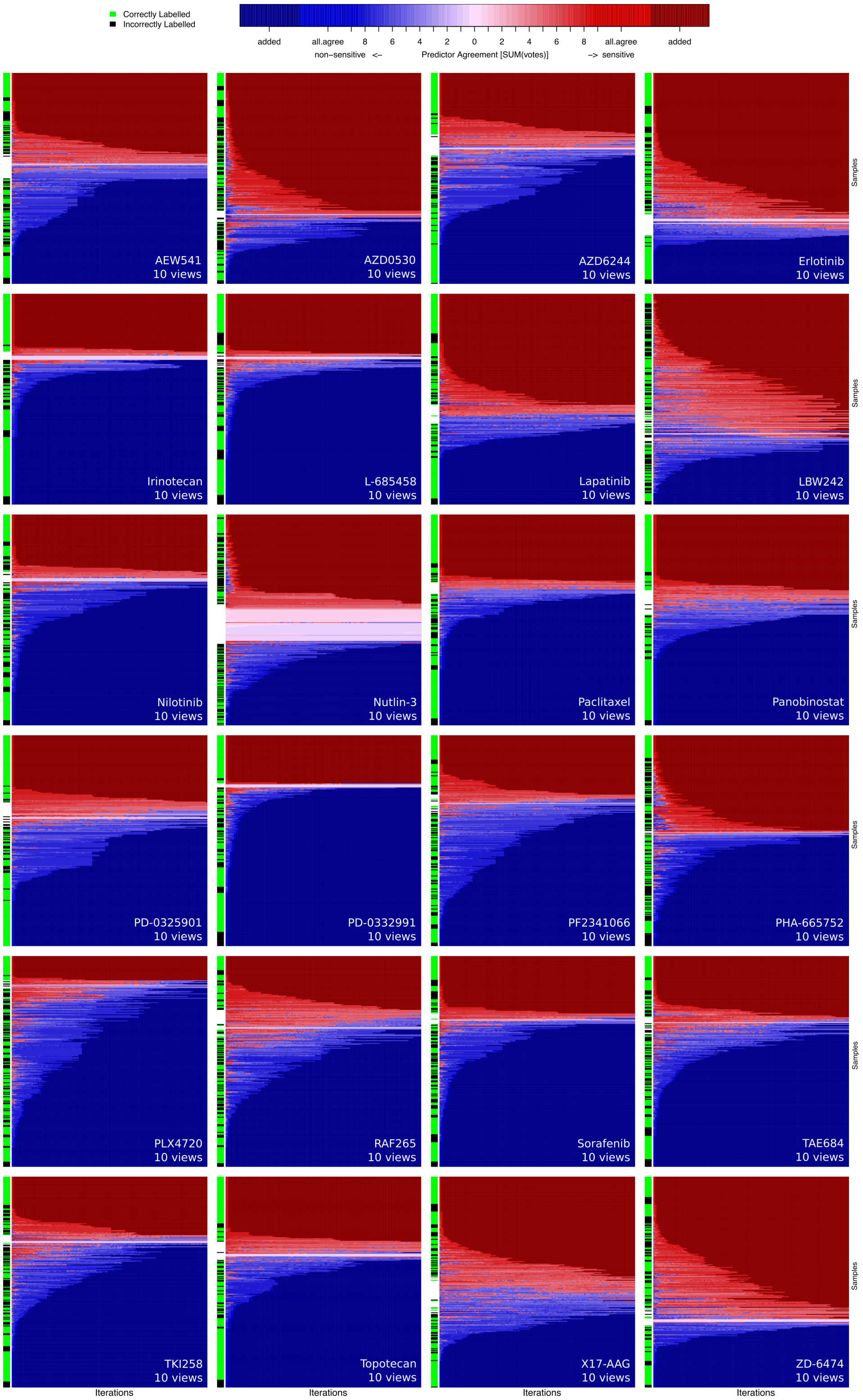

Fig. S6. Visualization of the label learning validation showing confidence of predictions for each unlabeled cell line at every iteration for all 24 CCLE drugs. Drug names in bottom right corner.

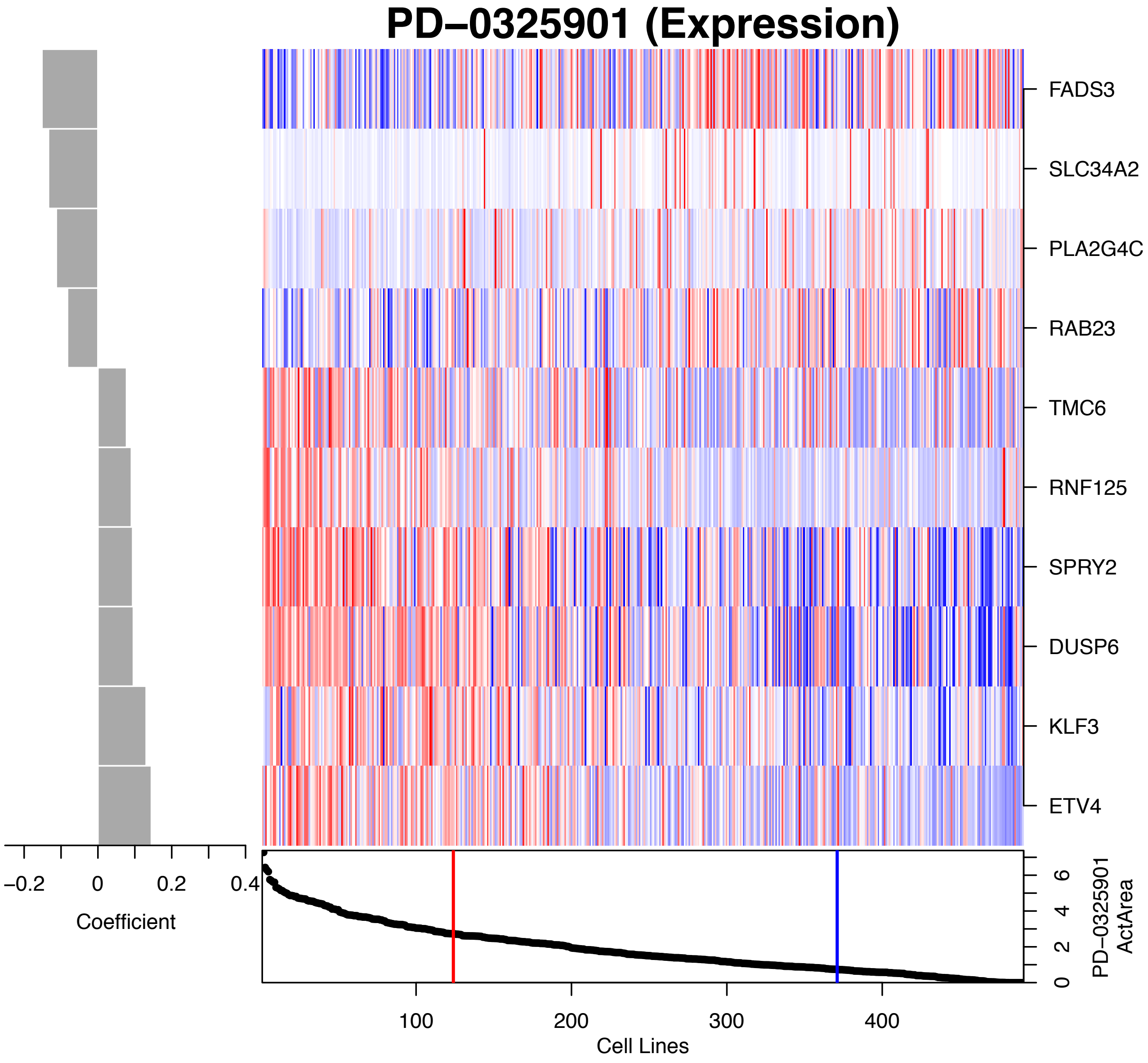

Fig. S7. Most important features in the PLATYPUS baseline expression view for PD-0325901 sensitivity prediction. The ten features with the highest absolute coefficient in the trained elastic net model are shown (left panel). The gene expression values are normalized and shown in continuous colors from blue (low expression) to red (high expression). The drug response values in the lower panel determine the sample sorting, but were used as binary labels as implied by the vertical lines: samples left of the red line are defined as sensitive, right of the blue line as non-sensitive.

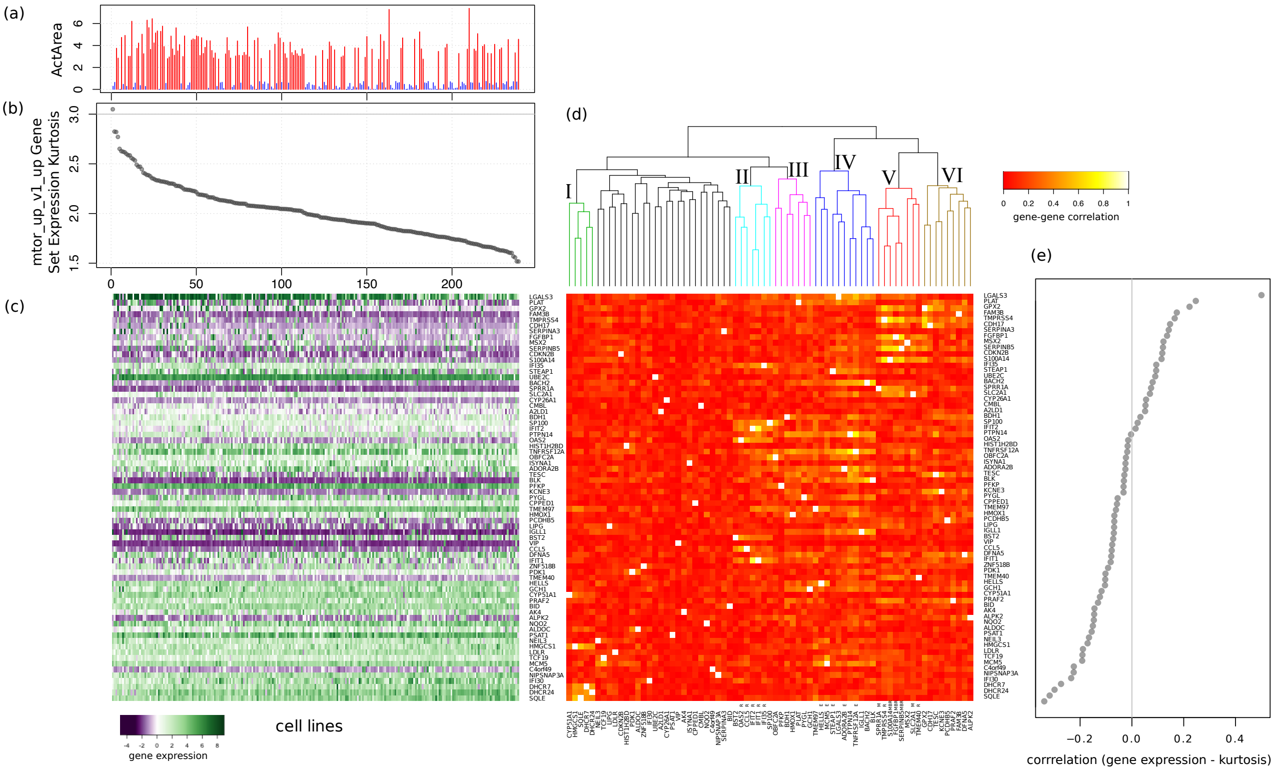

Fig. S8. MTOR_up_V1_up gene set feature and its relationship to expression and outcome. (a) ActArea for each cell line, sorted by the MTOR gene set kurtosis value; a higher proportion of resistant cells (red lines) are associated with higher kurtosis values of this gene set (left side) compared to sensitive cells (blue lines). (b) MTOR gene set kurtosis value for each cell line. (c) RNA Expression of the genes within the gene set. (d) Same genes as in (c), now showing gene–gene expression correlation. Tree shows hierarchical clustering of the genes and highlights groups of similar genes. Genes involved in EGFR signaling are marked with E, metastasis with M, basal vs mesenchymal BRCA with B, and resistance to several cancer drugs with R. (e) Correlation between gene expression and MTOR gene set kurtosis value determines sorting of genes in (c) and (d).

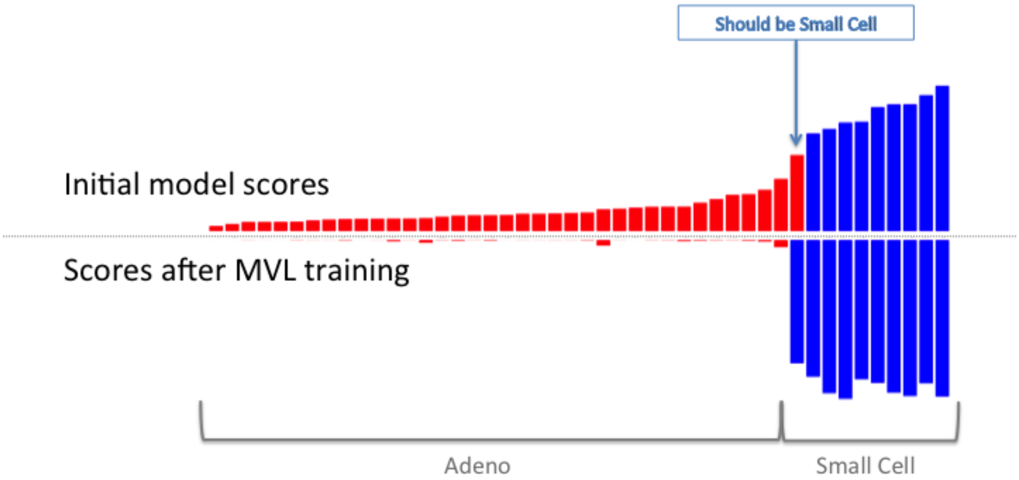

Fig. S9. Scores for the training set labels before and after PLATYPUS label learning.

Supplemental References

1. R. Aggarwal, J. Huang, J. J. Alumkal, L. Zhang, F. Y. Feng, G. V. Thomas, A. S. Weinstein, V. Friedl, C. Zhang, O. N. Witte *et al.*, *Journal of Clinical Oncology* **36**, 2492 (2018).
2. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin *et al.*, *Nature* **483**, p. 603 (2012).
3. A. Liaw and M. Wiener, *R News* **2**, 18 (2002).
4. J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, P. Hintsanen, S. A. Khan, J.-P. Mpindi *et al.*, *Nature biotechnology* **32**, p. 1202 (2014).
5. J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren, *Nature* **485**, 376 (May 2012).
6. M. K. C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang and C. Candan., *caret: Classification and Regression Training*, (2016). R package version 6.0-64.
7. A. Golugula, G. Lee, S. R. Master, M. D. Feldman, J. E. Tomaszewski, D. W. Speicher and A. Madabhushi, *BMC bioinformatics* **12**, p. 483 (January 2011).
8. M. Griffith, O. L. Griffith, A. C. Coffman *et al.*, *Nature methods* , 1 (October 2013).
9. P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor *et al.*, *Nature methods* **10**, 221 (March 2013).
10. A. Sokolov, C. Funk, K. Graim, K. Verspoor and A. Ben-Hur, *BMC bioinformatics* **14 Suppl 3**, p. S10 (January 2013).
11. M. White, X. Zhang, D. Schuurmans and Y. liang Yu, Convex multi-view subspace learning, in *Advances in Neural Information Processing Systems 25*, eds. F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (Curran Associates, Inc., 2012) pp. 1673–1681.
12. W. E. Johnson, C. Li and A. Rabinovic, *Biostatistics* **8**, 118 (2007).
13. A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo and J. P. Mesirov, *Bioinformatics* **27**, 1739 (2011).
14. J. Friedman, T. Hastie and R. Tibshirani, *Journal of statistical software* **33**, p. 1 (2010).
15. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545 (2005).
16. M. Rask-Andersen, S. Masuram and H. B. Schiöth, *Annual review of pharmacology and toxicology* **54**, 9 (2014).
17. G. Eschenburg, A. Eggert, A. Schramm, H. N. Lode and P. Hundsdoerfer, *Cancer research* **72**, 2645 (2012).
18. Y.-M. Li, M. Xu, M.-T. Lai, Q. Huang, J. L. Castro, J. DiMuzio-Mower, T. Harrison, C. Lellis, A. Nadin, J. G. Neduvelil *et al.*, *Nature* **405**, 689 (2000).
19. L. T. Vassilev, B. T. Vu, B. Graves, D. Carvajal, F. Podlaski, Z. Filipovic, N. Kong, U. Kammlott, C. Lukacs, C. Klein *et al.*, *Science* **303**, 844 (2004).
20. A. V. Galkin, J. S. Melnick, S. Kim, T. L. Hood, N. Li, L. Li, G. Xia, R. Steensma, G. Chopiuk, J. Jiang *et al.*, *Proceedings of the National Academy of Sciences* **104**, 270 (2007).
21. J. Da Cunha, P. Galante, J. De Souza, R. De Souza, P. Carvalho, D. Ohara, R. Moura, S. Oba-Shinja, S. Marie, W. Silva *et al.*, *Proceedings of the National Academy of Sciences* **106**, 16752 (2009).
22. F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, L. Zhang, Y. Song, X. Liu, J. Zhang *et al.*, *Nucleic acids research* , p. gkr797 (2011).
23. G. S. Cowley, B. A. Weir, F. Vazquez, P. Tamayo, J. A. Scott, S. Rusin, A. East-Seletsky, L. D. Ali, W. F. Gerath, S. E. Pantel *et al.*, *Scientific data* **1** (2014).
24. X. Xie, J. Lu, E. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander and M. Kellis,

Nature **434**, 338 (2005).

25. V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer *et al.*, *Nucleic acids research* **34**, D108 (2006).
26. M. J. Alvarez, Y. Shen, F. M. Giorgi, A. Lachmann, B. B. Ding, B. H. Ye and A. Califano, *Nature genetics* **48**, p. 838 (2016).
27. L. E. MacConaill, C. D. Campbell, S. M. Kehoe, A. J. Bass, C. Hatton, L. Niu, M. Davis, K. Yao, M. Hanna, C. Mondal *et al.*, *PloS one* **4**, p. e7887 (2009).
28. E. Khurana, Y. Fu, J. Chen and M. Gerstein, *PLoS Comput Biol* **9**, p. e1002886 (2013).
29. Y. Yang, M. Wislez, N. Fujimoto, L. Prudkin, J. G. Izzo, F. Uno, L. Ji, A. E. Hanna, R. R. Langley, D. Liu *et al.*, *Molecular cancer therapeutics* **7**, 952 (2008).
30. C. D. Allis, S. L. Berger, J. Cote, S. Dent, T. Jenuwien, T. Kouzarides, L. Pillus, D. Reinberg, Y. Shi, R. Shiekhattar *et al.*, *Cell* **131**, 633 (2007).
31. V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu *et al.*, *Nucleic acids research* **42**, D1091 (2014).
32. D. Keppler, Multidrug resistance proteins (mrps, abccs): importance for pathophysiology and drug therapy, in *Drug Transporters*, (Springer, 2011) pp. 299–323.
33. F. P. Guengerich, *Chemical research in toxicology* **21**, 70 (2007).
34. K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov *et al.*, *Cell* **158**, 929 (2014).
35. HUGO Gene Nomenclature Committee (HGNC), Gene family: Chromatin-modifying enzymes Accessed July 15, 2015.