Evaluating the consistency of large-scale pharmacogenomic studies

Raziur Rahman, Saugato Rahman Dhruba, Kevin Matlock, Carlos De-Niz, Souparno Ghosh and Ranadip Pal*

*Corresponding author: Ranadip Pal, Department of Electrical and Computer Engineering, 1012 Boston Avenue, Lubbock, TX 79409, USA.
E-mail: ranadip.pal@ttu.edu, or ranadip@gmail.com

Abstract

Recent years have seen an increase in the availability of pharmacogenomic databases such as Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE) that provide genomic and functional characterization information for multiple cell lines. Studies have alluded to the fact that specific characterizations may be inconsistent between different databases. Analysis of the potential discrepancies in the different databases is highly significant, as these sources are frequently used to analyze and validate methodologies for personalized cancer therapies. In this article, we review the recent developments in investigating the correspondence between different pharmacogenomics databases and discuss the potential factors that require attention when incorporating these sources in any modeling analysis. Furthermore, we explored the consistency among these databases using copulas that can capture nonlinear dependencies between two sets of data.

Key words: pharmacogenomic databases; database dependencies; copulas; pairwise relationships

Background

A pharmacogenomic system is composed of numerous genes that creates a complex response to drug application [1]. Pharmacogenomics is a comparatively new field in medical science that integrates Pharmacology (the science of drugs) and Genomics (the study of genes and their functions) to invent effective, nontoxic medications that can be customized to a person's genetic makeup. The majority of current therapy selections are population-based and thus fails to recognize the specific effect of an individual patient's genetic factors. After the discovery of the human genome [2, 3] and subsequent pharmacogenomics studies, the effect of a drug on an

individual's genetic makeup is becoming more visible, leading to different approaches that can more accurately predict the effects of medication on a patient. This approach, termed personalized medicine, can provide medical care to heterogeneous health problems [4] such as cardiovascular disease [5], Alzheimer's disease, Cancer [6], HIV/AIDS and asthma.

A diverse group of cell lines (CLs) belonging to a disease can be used to obtain biological characterizations such as from the genome [copy number variations (CNV), single-nucleotide polymorphisms (SNP), exome sequencing], epigenome (histone modification, DNA methylation), transcriptome [RNA sequencing (RNA-seq), microarray gene expression], proteome [reverse

Raziur Rahman is a doctoral student in the Department of Electrical and Computer Engineering at Texas Tech University. His research topics are related to application of machine learning in precision medicine.

Saugato Rahman Dhruba is a doctoral student in the Department of Electrical and Computer Engineering at Texas Tech University. He is currently working on application of transfer learning techniques in bioinformatics.

Kevin Matlock is a doctoral student in the Department of Electrical and Computer Engineering at Texas Tech University. He works in high-throughput data analysis for computational biology.

Carlos De-Niz is a doctoral candidate in the department of Electrical and Computer Engineering at Texas Tech University. His research is related to RNA sequencing and data science.

Souparno Ghosh is an assistant professor in the Department of Mathematics and Statistics at Texas Tech University. His research group focuses on statistical bioinformatics.

Ranadip Pal is an associate Professor in the Department of Electrical and Computer Engineering at Texas Tech University. His research group focuses on computational biology for precision medicine.

Submitted: 2 March 2018; **Received (in revised form):** 4 May 2018

© The Author(s) 2018. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

phase protein array (RPPA), liquid chromatography–mass spectrometry] or metabolome (nuclear magnetic resonance) with each level having a part in the translation of information coded in DNA to functional activities in the cell [7]. *In vivo* and *in vitro* pharmacological sensitivity studies using small-molecule components across panels of molecularly characterized cancer cell lines (CCLs) have assisted to understand the cellular activity of many compounds and assign mechanisms of drug actions [8]. In this relevant context of drug sensitivity prediction for personalized therapy, a number of machine learning models such as elastic net (EN), univariate and multivariate regression model, decision trees, neural nets and *k*-nearest neighbors (*k*-NN) have been proposed [9–11].

Standardization of pharmacological databases is a major concern for bioinformaticians for comparison and quality assurance purposes. Two recently studied projects on pharmacogenomics, the Cancer Cell Line Encyclopedia (CCLE) and the Genomic of Drug Sensitivity in Cancer (GDSC), were published with high hopes among researchers that these would allow for big data analyses of genomic characterization profiles and drug sensitivity. However, a subsequent meta-analysis [12] observed inconsistencies among the studies and questions were raised as to the credibility of the generated data.

This article is divided into two parts. In the first part, we review recent progress in the field of pharmacological studies along with the major arguments in favor of the consistency or inconsistency of these studies. Afterward, we explore the consistency of studies from two perspectives, biomarker selection and linear relationship. In the second part, we analyze the CCLE [13] and GDSC [14] studies from a new perspective. Until now, the consistency of databases was evaluated by directly comparing the responses of a single drug. But because of differences in experimental protocols, finding high concordance was always difficult. Instead, we will compare the responses for pairs of drugs in a single study and then investigate whether the structure of said responses is consistent with the second study. We use copulas to study the dependency structure between drug and CL pairs among databases and illustrate that the pair-wise dependency is maintained to a large extent. To the best of our knowledge, this is the first time that these major pharmacogenomics studies have been compared from the drug pair aspect. This article also includes comparisons of CCLE and GDSC with another popular repository, the NCI-60 Human Tumor Cell Lines Screen [15].

Large pharmacogenomic studies

In recent years, a large number of high-throughput studies have been performed on CCLs to investigate the effects of cancer drugs on a specific cancer type. These results are then stored in a pharmacogenomic databases for further community analysis. However, these studies differ in the protocols they followed associated with pharmacological assays and cell viability drug screenings. This section will provide an overview of the protocols followed by several major studies along with the genomic and functional characterization data developed through these processes.

The CCLE is a joint collaborative work by the Broad Institute, the Novartis Institutes for Biomedical Research and the Genomics Institute of the Novartis Research Foundation to develop a genetic profile of 947 human CCL from 36 distinct tumor types [13]. In addition, they have tested the sensitivity of 24 anticancer compounds on ~500 of the CCL. The technology platforms used to characterize the CCL include the Affymetrix U133 plus 2.0 arrays to calculate gene or messenger RNA expression

[16], CL transcriptomic sequence by RNA-seq, high-density SNP arrays (Affymetrix SNP 6.0) for the DNA copy numbers [17] and mutation information are calculated using next-generation sequencing of > 1600 genes and high-throughput genotyping platform (OncoMap). To measure drug sensitivity, CCLE has generated eight-point dose–response curves using logistical sigmoidal function fitting [13, 18]. These curves are then used to calculate different drug sensitivity metrics, including IC_{50} (the concentration of the compound that provides 50% inhibition of the CL), EC_{50} (the concentration that provides half the maximum inhibition of the compound), A_{max} (maximal effect level of compound) and AUC (area under the percent viability curves) [19].

The Cancer Therapeutics Response Portal (CTRP) [20] has used the genomic characterizations data from the CCLE project [13] and then generated responses of 242 CCL to 354 small-molecular compounds (including 35 FDA-approved drugs, 54 drugs in clinical trials and 265 probes), which target particular nodes of significant cellular processes. Each CCL was grown, plated and then treated with each compound at eight different concentration levels for 72 h. CellTiter-Glo (CTG) was used for assaying the sensitivity by calculating cellular ATP levels; this serves as a surrogate for cell number and growth measurements. However, CTRP only provides the computed AUC as the sensitivity measurement [20]. In a recent update, CTRP has increased the number of CCL to 860 and the number of targeted drugs to 461 [21].

The GDSC [22] database was generated through a collaboration by the Cancer Genome Project (CGP) [14] and the Center for Molecular Therapeutics at Massachusetts General Hospital. They have genomically characterized >1000 different CCL from 29 distinct tumor types. Characterization of the CCL includes information of somatic mutations in 75 cancer genes, genome-wide gene copy number analysis for amplification and deletion, targeted screening for seven types of gene rearrangements, markers of microsatellite instability, tissue type and transcription data. The genomic data sets contained within GDSC have been collected from the Catalogue of Somatic Mutations in Cancer (COSMIC) [23] database. In the original database, 138 anticancer therapeutic compounds, including both targeted and cytotoxic drugs, have been screened for 329 to 668 CCL per drug, giving a total number of 73 169 CL–drug interaction measurements. Cell viability is measured using fluorescence-based cellular assays 72 h post-drug treatment. For each compound, nine concentrations have been tested and sensitivity metrics are given as AUC and IC_{50} values. In the current version of the database (GDSC v6), they have increased the number of tested compounds to 265 anticancer drugs [24]. The database contains the gene expression, mutation and CNV information of >23 000 genes. In addition, Iorio *et al.* have generated an oncogenomic alternations map in human tumors using data from The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) and other sources. This map consists of the mutation pattern of cancer genes, focal recurrently aberrant copy number segments from SNP6 array profiles and gene promoters (iCpGs) from DNA methylation data [24].

The NCI-60 [15] data set has used complementary DNA microarrays to detect a variation in 8000 genes among 60 CCLs. Other genomic information found in this data set includes CNV, mutation, mRNA, microRNA (miRNA), DNA methylation and protein expression. In 2013, a collaboration between NCI and the Dialogue on Reverse Engineering Assessment and Methods (DREAM) [9] created a database where six genomic characterizations (gene expression, methylation, RNA sequencing, whole-exome sequencing, RPPA and CNV) are included for 53 breast CCL along with sensitivity measurements for 35 anticancer drugs. An

extension of this database, the GRAY database [25] has profiled the CNV, mutations, gene and isoform expression, promoter methylation and protein expression of 70 breast CCLs. In addition, they have given sensitivity information in the form of GI_{50} (concentration at which 50% growth inhibition is achieved) for 90 anticancer compounds, 18 of which are FDA-approved drugs.

A large numbers of human patient tumors have been profiled and assayed in TCGA [26, 27] to discover the molecular aberrations among genes using proteomic and epigenetic expression. However, this data set has yet to include sensitivity measurements for anticancer drugs. Among the data categories available in TCGA, their data portal provides RPPA, DNA methylation, CNV, mutation, miRNA and gene expression for a total of 5074 tumor samples.

Genentech Cell Line Screening Initiative (gCSI) [28] has reported on 16 anticancer drugs applied to 410 CCLs. Sensitivity measurements provided are the mean of the fitted viability curve (equivalent to AUC) and IC_{50} values. An extension of the gCSI database is the Genentech (GNE) database [29], which has provided the RNA sequencing and SNP array analysis for 675 human CCL along with the responses to the drug pictilisib (PI3K inhibitor) and cobimetinib (MEK inhibitor).

A new data set of drug responses was profiled by the Institute for Molecular Medicine Finland (FIMM) compound testing assay [30, 31], covering 308 drugs across 106 CCLs using CTG to measure CL viability. In the Personal Genome Project (PGP) [32], mutation and gene expression data have been profiled for 87 CCL extracted from lung, breast and colorectal tumors. For these CLs, the IC_{50} values for the Aurora kinase inhibitor PF-03814735 are given. A database provided by Harvard Medical School [33] gives RPPA measurements of 17 signaling proteins and 4 cell state markers in 10 CLs. In addition, they also provide apoptosis and cell viability values of five drugs measured six time points for seven separate concentrations.

The Library of Integrated Network-based Cellular Signatures (LINCS) [34] project is an NIH-funded program, where currently

six centers generate data, which are Drug Toxicity Signature Generation Center, HMS LINCS Center, LINCS Center for Transcriptomics, LINCS Proteomic Characterization Center for Signaling and Epigenetics, MEP LINCS Center and NeuroLINCS Center. The entire data portal consists of 387 data sets, 41 847 small molecules, 1127 CLs and 978 genes. These data sets are based on transcriptomics, binding, imaging, proteomics and epigenomics studies.

The ICGC [35] is a collaborative effort to gather large-scale cancer genome studies. In ICGC, gene expression, copy number alterations, simple somatic mutations, structural rearrangements, exon junctions, miRNAs and DNA methylation databases [36] of 50 cancer subtypes by systematically studying >25 000 cancer genomes are publicly available. Till now, ICGC has committed to 90 CGPs, including TCGA, Tumor Sequencing Project (TSP) and Sanger Cancer Genome project, and publishes data from time to time. The last release (v26) included 57 K mutated gene information for 21 cancer primary sites. The common cancer genome studies and consortium are summarized in a Tabular form in Table 1.

Current drug databases

This section presents a short review for a selected few of the current drug databases such as DrugBank, Drug-Gene Interaction Database (DGIdb), SuperDRUG2, DCDB, TTD, PharmaGKB and STICH that combine different pharmacological studies under the same roof. Note that, we have focused on the drug-specific databases here, and therefore, descriptions of large popular repositories (such as KEGG, PubChem and HMDB) are not provided.

DrugBank [37] is a freely accessible database first launched in 2006 and contains comprehensive molecular information about both approved and experimental drugs, drug mechanisms, interactions and biological targets (i.e. sequence, structure and pathway). The latest release (version 5.0.11) holds 11 033 drug entries with 2521 approved small-molecule drugs, 949 approved biotech (protein/peptide) drugs, 111 nutraceuticals

Table 1. Selected information about some Cancer Genome Studies and Consortium

Database name	Participating institute	Sensitivity assay	Genomic characterizations data sets	Experiment on number of CLs	Experiment on number of drugs	URL
CCLE	Broad Institute, Novartis Institutes for Biomedical Research and Genomics Institute of the Novartis Research Foundation	Affymetrix U133 plus 2.0 arrays	Gene Expression, SNP6, Mutation	947	24	http://portals.broadinstitute.org
GDSC	CGP and Center for Molecular Therapeutics at Massachusetts General Hospital	Affymetrix U133A arrays	Gene expression, Mutation, CNV	>1000	265	http://www.cancerxgene.org
CTRP	Center for the Science of Therapeutics at the Broad Institute	CTG	Same as CCLE	860	461	https://portals.broadinstitute.org/ctrp/
NCI60	National Cancer Institute (NCI)	five-dose assay	CNV, Mutation, Protein Expression	60	>52 000	https://dtp.cancer.gov/discovery_development/nci-60/
LINCS	Drug Toxicity Signature Generation Center, HMS LINCS Center, LINCS Center for Transcriptomics, LINCS Proteomic Characterization Center for Signaling and Epigenetics, MEP LINCS Center and Neuro LINCS Center	Imaging Assays	Transcriptomics, Binding, Imaging, Proteomics and Epigenomics	1127	41847	http://lincsportal.ccs.miami.edu/dcic-portal/

and over 5112 phase I/II/III investigational drugs, with each entry being linked to sequences of 4911 nonredundant proteins (i.e. drug targets, enzymes, transporters and carriers). DrugBank 5.0 has also added novel information for hundreds of drugs on different pharmacology levels, i.e. pharmacometabolomic, pharmacotranscriptomic or pharmacoproteomic data sets. It also boasts major improvements of different existing tools and data formats such as the spectral viewing and spectral search tools, spectral data formats, chemical taxonomies, chemical ontologies and text and structure searching/matching.

SuperDRUG2 [38] is an updated version of the conformational drug database, SuperDRUG [39], that contains comprehensive information for 4587 approved and marketed drugs in two categories: small molecules (3982 drugs) and biological/other drugs (605 drugs). The database is intended to serve as a 'one-stop resource' containing data from multiple widely distributed sources to provide information on drug chemical structures (2D and 3D), dosage, regulatory details, biological targets, physicochemical properties, external identifiers, side effects, pharmacokinetics and so on. It provides multiple search options to facilitate analyses such as 2D/3D structural similarity calculation, potential drug-drug interaction identification in complex drug regimens and so on. The database also provides additional features to simulate the plasma concentration versus time curves from pharmacokinetic data and 3D superposition to superimpose drugs of interest with ligands with known targets in 3D structures.

The DGIdb [40] gathers and catalogs drug interaction data for collections of altered genes (via mutation or otherwise) from

multiple sources along with the gene druggability information from pathway memberships, molecular functions and gene families from the Gene Ontology (GO), dGene and druggable genome lists [41]. The latest DGIdb release (version 3.0) exhibits a major update in terms of data sources, volume and usability (API features), resulting in 56 309 interaction claims from 30 sources with a substantial expansion of the druggable gene catalogs.

The Drug Combination Database (DCDB) [42] claims to be the first available database to collect and organize information on drug combinations with the aim of facilitating novel systems-oriented drug discovery. The current version of DCDB (v2.0) has 1363 drug combinations available in three categories—approved (330), investigational (phase I/II/III/IV trials, 1033) and unsuccessful (237) for 904 individual drugs and 805 targets. DCDB also contains comprehensive information for each type, i.e. for drug combinations, it provides the combined activity/indications, possible mechanism, component interactions and development status; for individual drugs, the chemical, pharmacological, pharmaceutical properties and known targets; and for each drug target, its sequence, function annotation and affiliated pathway.

Table 2 provides overview of several drug databases.

Bioinformatics tools

This section presents a short review of the Bioinformatics tools available for analyses of Pharmacogenomic databases. Numerous online tools are available in this regard such as

Table 2. Selected information about some current drug databases

Database name	Current release	Description	URL
DrugBank [37]	v5.0.11	A comprehensive online database containing drug molecular data, mechanisms, interactions and biological targets. DrugBank has 11 033 drug entries with 2521 small molecules, 949 biotech drugs, 111 nutraceuticals and over 5112 experimental drugs. Each entry is linked to 4912 nonredundant protein sequences (i.e. drug target/enzyme/transporter/carrier)	http://www.drugbank.ca
SuperDRUG2 [38]	v2.0	A conformational drug database providing comprehensive information for 4587 approved/marketed drugs (3982 small molecules and 605 biological/other drugs). Drug annotation provides data on 2D/3D chemical structures, dosage, regulatory details, targets, physicochemical properties, external identifiers, side effects, pharmacokinetics and drug-drug interactions	http://cheminfo.charite.de/superdrug2/
DGIdb [40]	v3.0	An interaction database assembling 56 039 drug-gene interaction claims for mutated or otherwise altered gene lists implicated in diseases. DGIdb also catalogs the gene druggability information based on factors such as pathway memberships, molecular functions and gene families	http://www.dgiddb.org
DCDB [42]	v2.0	The first available drug combination repository to assemble comprehensive information for 1363 drug combinations (330 approved, 1033 investigational and 237 unsuccessful) for 904 individual drugs and 805 drug targets. DCDB provides information for each data type (i.e. drug combinations, drugs or targets) in details.	http://www.cls.zju.edu.cn/dcdb/
Pharmacogenomics Knowledgebase (PharmGKB) [43]	–	A pharmacogenomics-based database that curates pharmacogenetics information to identify gene-drug associations and genotype-phenotype relationships. In current version, information of 20 017 genetic variants (including SNPs, haplotypes, CNVs and indels), 3753 clinical annotations, 65 important pharmacogenes, 130 pathways and 641 drugs with 498 drug labels is available	https://www.pharmgkb.org/
DrugCentral [44]	v9.4	An online drug compendium where chemical entities, pharmaceutical products, drug mode of action, indications, pharmacologic action of 4509 active ingredients are annotated, which are FDA, EMA and PMDA approved. The properties of the drugs are aggregated from ChEMBL, KEGG, ATC, EPC and similar databases	http://drugcentral.org/

LINCS tools, COSMIC tools, STRING-DB, STITCH, Drug-Set Enrichment Analysis (DSEA), ChemMine tools, UniProtKB, PubChem tools and so on. We have described a selected few below. Table 3 also provides brief descriptions of a few of such bioinformatics tools.

The COSMIC [45] is a part of CGP from the Wellcome Sanger Institute in the UK and world's largest resource for expert-curated somatic mutation data for human cancers. The latest release of COSMIC, i.e. v84 (February 2018) contains over 5.5 million coding mutations combining genome-wide sequencing results from 33 291 tumors with manual curation of 25 807 papers across all cancers, as well as data for 18 million noncoding mutations, 18 926 gene fusions, 1.2 million abnormal copy number variants, 10 million abnormal expression variants and 8 million differentially methylated CpG dinucleotides. The website presents numerous dedicated tools for data exploration, including Cancer browser (provides disease-specific perspective via mutation exploration), Genome browser (provides genome-wide perspective to cancer genomics), Gene pages (summary of a specific gene data), Fusion genes, Drug resistance data, Hallmarks of Cancer, CONAN (copy number analysis tool) and so on.

The Search Tool for Retrieval of Interacting Genes/Proteins (STRING) [46] is a database containing the known and predicted

protein interactions—both direct (physical) and indirect (functional), deriving from computational prediction, knowledge transfer between organisms and interaction data from other sources. The latest release, STRING v10.5, covers 1.4 billion interactions (with individual confidence scores) for 9.6 million proteins from 2301 organisms. Multiple ways are available to access the protein networks, including the Web interface for small- to medium-scale networks and programmatic access for large-scale network analyses via a REST-API, a R/Bioconductor package and the data payload mechanism through adding [supplementary data](#) to the website (i.e. user-provided interactions and protein-centric information) [46]. Besides, users are provided with statistical analysis results for each network via necessary alerts for various criteria, which is particularly useful in the case of functional characterization of multiple protein sets. Additionally, STRING has also developed an App for the Cytoscape software framework allowing for easy retrieval, visualization and analyses of networks of hundreds to thousands of proteins via a GUI using protein names, diseases or PubMed queries.

ChemMine Web Tools [47] is an online service for analysis and clustering of small molecules by structural similarities, physicochemical properties or custom data types [47]. It provides a Web interface to a set of cheminformatics and data

Table 3. Selected information about some bioinformatics tools

Tool	Current release	Description	URL
LINCS Tools [48]	–	There are number of LINCS tools available that help users analyzing LINCS data sets. Web and software platforms such as L1000CDS ² , iLINCS, Drug-Pathway Browser, Drug/Cell-line Browser, Enricher, etc., analyze features of LINCS data sets such as expression profiles, signatures, drug–target, pathway, CLs, responses and so on in versatile ways	http://www.lincsproject.org/LINCS/tools
COSMIC [45]	v84.0	COSMIC presents several dedicated tools for data exploration, including Genome browser, Gene pages, Cancer browser, Fusion genes, Drug resistance data, Hallmarks of Cancer, COSMIC-3D, Cancer Gene Census, Mutation signatures, CONAN	http://cancer.sanger.ac.uk
STRING [46]	v10.5	STRING contains both known and predicted protein interaction networks. Small- to medium-scale networks are available via Web interface, while large-scale networks are analyzed via R/Bioconductor package, REST-API or data payload mechanism by adding supplementary data along with statistical analysis results. Additionally, a Cytoscape-based app is available for easy retrieval, visualization and analysis of protein networks via GUI	https://stringdb.org
ChemMine Web Tools [47]	–	ChemMine is a Web-based service for analysis and clustering of small molecules that provides an interface to a set of cheminformatics and data mining tools. Compounds are imported to workbench by drawing, copy/paste, from local files or PubChem search. Functionalities include in data visualization, structure comparisons, similarity searching, compound clustering and chemical property prediction	http://chemmine.ucr.edu/
CLUE [49]	v1.1	CLUE is a software platform that performs chemical and genetic perturbation analysis in addition to providing over 1 million expression profiles. It helps users to integrate latest versions of high-dimensional perturbation data sets, which are from multiple assays, cell types and different dose and treatment conditions and then facilitates interoperability and implements Web applications (connectivity among perturbagens, gene expression signatures, protein sets, etc.) with GUI	https://clue.io/
DSEA [50]	v1	DSEA identifies phenotype-specific pathways that are targeted by majority of the drugs in a set based on drug-induced gene expression profiles. It follows the same algorithm of GSEA but with an inverse preparation and interpretation of the data. DSEA gives more weights to the pathways that are most dysregulated in the set of selected drugs in comparison with the full set of drugs in the database	http://dsea.tigem.it

mining tools useful for various chemical genomics and drug discovery analyses along with a programmatic access via the R library 'ChemmineR'. Compounds can be imported to the workbench by drawing, copy/paste from local files or from a PubChem search including an online molecular editor. ChemMine Tools provide functionalities in five major categories, i.e. data visualization, structure comparisons, similarity searching, compound clustering and chemical property prediction. ChemMine similarity toolbox uses two algorithms—atom pairs as descriptors and Tanimoto coefficients as similarity measures or users can choose the similarity coefficients to identify the maximum common substructure shared. For efficient data mining for chemical structure and bioactivity space, ChemMine Tools service provides similarity search methods with PubChem interfaces via its data exchange feature. It also provides an interface to the property prediction module of the JOELib package that can calculate 38 physicochemical property values. The resulting property tables can be further processed by sending them to the Clustering toolbox that uses either hierarchical clustering, multidimensional scaling or binning clustering algorithm.

Arguments in favor of the inconsistency of pharmacological databases

The early gene expression studies two decades ago were fraught with noisy measurements resulting in discrepancies in similar studies [51, 52]. Extensive research conducted in subsequent years to standardize data collection and analysis approaches [53, 54] and the development of robust expression-level measuring platforms [55] minimized the inconsistencies in gene expression observations from separate studies. In recent pharmacogenomic studies, researchers have noticed concordance between the collected gene expressions for different databases. For instance, when Haibe-Kains et al. [12] compared the gene expression profiles between 471 common CLs of the CCLE and GDSC, they observed a median Spearman correlation of 0.85. However, other studies [12, 56–59] found that the pharmacological drug responses in the two databases were observed to be discordant. In the following section, we highlight the points that support the inconsistency of drug responses.

Observed inconsistencies in direct comparison of drug responses

Haibe-Kains et al. [12] considered comparison between drug sensitivity measures of 15 common drugs of CCLE and GDSC databases and observed median correlation of 0.28 and 0.35 for drug response metrics IC_{50} and AUC, respectively. However, a direct comparison may not be fair, as the two studies followed different protocols, such as the range of drug concentrations that have been tested. Filtering out the insensitive CLs did not result in much higher consistency, as only a couple of drugs exceeded a correlation of 0.5. Furthermore, discrete classification of CLs as resistant, intermediate and sensitive using the Waterfall method [13] did not increase the concordance either as measured by Cohen's κ coefficient [60].

Another important factor, which could potentially be the cause behind the inconsistency between CCLE and GDSC, is the method used to fit the dose–response curves and summarizing them for calculating IC_{50} or AUC statistics [59]. For instance, GDSC used the Bayesian sigmoid method [22] with extrapolation, while CCLE used the maximum concentration tested [13] for IC_{50} estimation. In CCLE, there is a robust cooperation

between Compound and Response Summary indicating that both the compound and thereafter the potential to outline the compound's dose response measurements have viable effect on the model performance, while in GDSC data set, Response Summary does not play that much effect. Cell population heterogeneity and cell-to-cell viability in drug response can have effect in the structure of the dose–response curves, which could result in shallow curves with reduced edges or Hill slope [61]. It has been suggested that, instead of directly comparing the IC_{50} or AUC, if raw dose response data were compared, then inconsistency factors could have been investigated further [59]. Unfortunately, such data are not currently available for both studies.

Instead of directly comparing IC_{50} or AUC, Safikhani et al. [56] introduced two metrics, the area between two dose–response curves (ABC) and Matthews correlation coefficient (MCC) to estimate the consistency of continuous and discrete drug sensitivities, respectively. Using these statistics, moderate consistency of drug responses has been found for a couple of drugs, but there are no statistics that have shown consistency across all the common drugs between CCLE and GDSC.

Observed inconsistencies while comparing genomic predictors of drug response

One of the important objectives of CCLE and GDSC studies were to identify genomic predictors of drug response. Haibe-Kains et al. [12] estimated gene–drug associations using a linear regression model fitting approach with gene expression as the predictor of drug sensitivity but observed low concordance between the studies (highest correlations for IC_{50} and AUC are 0.38 and 0.46, respectively). They observed that the overall correlation can be improved by using genes that are related to drug responses, but the results are still not satisfactory.

Another way to find the association of genes with drug sensitivity is by computing normalized enrichment scores with the help of over-represented GO terms [62, 63] from gene set enrichment analysis (GSEA) [64]. The overall correlation for drugs using GSEA enrichment scores was poor [12] except for a couple of cases where moderate correlation was found such as for the drugs AZD6244 and PD0325901. The correlation can however be increased by considering significantly enriched GO classes.

To check the possibility of predicting outcomes on an independent data set, Papillon-Cavanagh et al. [58] built genomic predictors from GDSC data set and validated those using the CCLE data set. Five linear methods consisting of both univariate and multivariate models were used to build genomic predictors. In nine of the drugs contained within GDSC, good performance was observed in terms of prediction using the selected genomic predictors and 10-fold cross-validation. But when these same predictors were used for the validation of both common and new CLs of CCLE (compared with GDSC), only a couple of drugs showed satisfactory performance. Using discordancy partitioning models, Rao et al. [65] have found reproducible biomarkers for every common drug of CCLE and GDSC, which other methods have failed to achieve.

Table 4 summarizes the various studies, which support the inconsistency of the large pharmacogenomic studies.

Arguments in favor of the consistency of pharmacological databases

Following the reports on the inconsistencies of drug sensitivities between CCLE and GDSC databases, attempts have been

Table 4. For showing inconsistency of CCLE and GDSC data sets, key factors like aspects or, methods or, sources discussed by different papers have been summarized here

Publication title	Key factors discussed	Aspect	Method used	Source of inconsistency
Inconsistency of large pharmacogenomic studies [12]	(i) Direct comparison (ii) Gene–drug associations (iii) Mutation (iv) Pathway-based correlations	IC ₅₀ , AUC, Biomarkers	Spearman correlation, Waterfall method, GSEA	Experimental protocol, drug sensitivity measurement
Revisiting inconsistency in large pharmacogenomic studies [56]	(i) Comparison methods (ii) Distribution of drug responses (iii) Drug targets	IC ₅₀ , AUC, Biomarkers	Pearson correlation, area between drug dose–response curves (ABC), MCC, Somers' Dxy rank correlation Cramer's V	Experimental protocol
Enhancing reproducibility in cancer drug screenings: how we move forward? [61]	(i) Experimental setup	IC ₅₀ , AUC, Hill Slope	–	Experimental protocol
Systematic assessment of analytical methods for drug sensitivity prediction from CCL data [59]	(i) Relation between compound and response summary	IC ₅₀ , AUC	Nonlinear models, ANOVA	Dose–response curve fitting
Comparison and validation of genomic predictors for anticancer drug sensitivity [58]	(i) Genomic predictor	Biomarkers	Linear univariate and multivariate models, concordance index	–

made to explain the discrepancies and arrive at approaches to correctly interpret the analytical output of large-scale pharmacogenomic studies. In this section, we discuss the primary factors noted by researchers to explain the discrepancies.

Biological variation among the methods used for data generation

This subsection considers the issues related to differences arising because of biological factors in the two databases such as assays measuring different states of the biological system or changes happening because of alterations in the CLs being measured.

Pharmacological assays

The inconsistencies among pharmacological databases could be a result of the difference in biological properties of the pharmacological assays, gene expression profiles, computational algorithms or any combination of the aforementioned factors [61, 66, 67]. Surprisingly, the gene expression profiles used in CCLE and GDSC that were obtained from microarray studies were highly concordant, whereas the pharmacological assays were different [66]. GDSC used the CellTiter 96 Aqueous One Solution Cell Proliferation Assay [68] from Promega as the pharmacological assay that measures a reductase-enzyme product after 72-h drug incubation as a measure of metabolic activity. On the other hand, CCLE used the CTG assay [69], also from Promega, as the pharmacological assay that uses the levels of ATP after 72–84 h of drug incubation as a measure of metabolic activity. The two assays used in the two databases are providing indices of the drug activity against the cells in two different ways, which restricts the mirroring of the experiments. For example, better correlation has been observed between CCLE [13] and GlaxoSmithKline (GSK) [70] that uses the same assay [71] as compared with correlation between GDSC and GSK [12], whereas multiplexing approaches have been used to reproduce experimental results and decrease inter-assay variability that results in the increase of data set concordance [72].

Furthermore, drug sensitivity measurement is another likely source of discordance as can be proved by the fact that perfect median correlation is achieved when using identical drug phenotypes with the actual gene expression data [12].

Recent studies [28, 66, 67] have pointed out to some of the factors that could influence the quantitative results obtained from such assays, such as different batches of fetal bovine serum (FBS) used for cell culture medium, cell seeding density, time and conditions of cell incubation before the drug is added, the coating on the plastic culture wells, intra-study batch or trend effects and other such obscure factors. For instance, when slow growing CLs are seeded at higher density, it causes the control wells to become confluent over the course of the assay and potentially constrains the growth of saturating CTG signals [28]. This will increase the mean viability, while the average drug sensitivity will decrease. The opposite phenomenon is observed for fast-growing CLs seeded at lower density. For example, for the drug PD0325901, increasing FBS has systematically increased mean viability [28].

To evaluate the relevance of methodology differences, Haverly et al. [28] reexamined all the 24 CLs and four drugs (PD0325901, erlotinib, lapatinib and paclitaxel) common in the three studies (CCLE, GDSC and gCSI) with CTG versus SYTO 60 fluorescent strain, fixed versus variable seeding and 5 versus 10% FBS. CL viability reduced by 3.6% when assessed using CTG as compared with SYTO 60 for the drug PD0325901, but no biases in the mean viability for the other three drugs were found [28]. Furthermore, compared with the GDSC SYTO 60 results, SYTO 60 values for broadly active drugs found by [28] were more congruent with CTG values from the primary gCSI and CCLE screens. The SYTO 60 assay has wider confidence interval than CTG increasing the variability and lowering the precision.

For further investigating on whether the biological properties are the main reason behind the inconsistency, Mpindi et al. [30] applied an experimental protocol similar to that used by CCLE to a new data set, FIMM [31]. FIMM and CCLE have the same readout (CTG assay [69]) and controls but different plate format (1536 versus 384 wells) and unstandardized cell

numbers. In addition, there was no standardization in the source, passage, cell media or the origin and handling of drugs. Despite these mismatches in the protocol, median correlation between CCLE and FIMM drug responses was high with a between-CLs correlation of 0.74. The median between-CLs correlation between GDSC and FIMM drug responses was 0.54 [30], potentially because of major differences in experimental protocol such as using the SYTO 60 fluorescent nucleic acid stain for the readout, lack of positive controls and plate formatting. In a recent study [28], gCSI observed that the laboratory-specific effects can potentially result in a greater bias compared with different readouts.

To evaluate the source of variation of high-throughput screening (HTS) data sets, Ding et al. [73] performed a study of inter- and intra-site experimental variability across skin CCLs treated with 120 different drugs screened separately in Sanford Burnham Prebys (SBP) Medical Discovery Institute and Translational Genomics Research Institute (TGen). Applying flexible linear regression modeling within an analysis of variance (ANOVA) context, it has been found that difference in laboratory protocols only explained 0.028% of the drug response variation, whereas plate (3.23%), drugs (45.5%), concentration (5.24%) and examined CLs (4.94%) explained nearly 60% of the variation. Further, ANOVA analyses on IC_{50} values stated in the GDSC and CCLE databases along with TGen and SBP data on the six common drugs and four CLs have shown that the laboratories were not substantial interpreters of IC_{50} values [73].

Predictive biomarkers of drug response

A key screening metric to show the consistency between the databases would be the identification of the same predictive biomarkers for drug responses. To identify predictive biomarkers, CCLE and GDSC both used the renowned penalized regression strategy of EN [74], which is effective in picking a small number of important molecular features out of thousands of candidate features. Haverty et al. [28] conducted a direct comparison between the EN results obtained from CCLE and a new database gCSI. Despite the possibility of picking equivalent or redundant features for different studies [75], gCSI and CCLE have revealed substantial similar results for most of the common drugs [28].

Another study has applied EN regression across 21 013 genomic features comprising expression, CNV and mutations and observed highly significant overlap of predictors for most of the drugs, even for drugs that have very few overlapping CLs [76]. For some of the drugs that had low correlation in drug sensitivities, better consistency was observed in terms of these biomarkers.

The biomarkers that have been selected by EN for one study have also performed well for univariate ridge regression [77] on the other study's response data [28, 76]. Using EN modeling on each database, [76] compared 4957 drug-gene associations and observed only one incongruent result between the two studies.

Stransky et al. [76] performed ANOVA test [78] using the overlapping CCLE and GDSC CLs to identify known genetic biomarkers of sensitivity or resistance. In at least one data set, the ANOVA-identified biomarkers were top molecular correlates for 13 of 15 compounds, whereas for both data sets, it was for 8 of 15 compounds. Furthermore, after fitting ANOVA to activity area, 14 drugs in GDSC and 15 drugs in CCLE showed consistency across data sets in terms of lineage-specific response associations [76]. Later Safikhani et al. [57] have raised questions regarding the results because of the use of same GDSC mutation

data across two databases and reusing CCLE genomic data for EN design in both data sets.

For consistency, Geeleher et al. [79] have suggested to only consider the target-positive CLs, which will also be sensitive to drugs. They have identified several instances where targeted agents are associated with drugs [79], such as BCR-ABL1 for nilotinib [80]; ERBB2 for lapatinib [81]; NQO1 expression for 17-AAG [82]; BRAF mutation for PD-0325901 [83], AZD6244 [84] and PLX4720 [85]; MDM2 for Nutin-3 [86]; and MET for Crizotinib [87]. However, a problem with considering target-positive CLs is that there will be significantly smaller number of CLs that can be compared [88] resulting in lowering the statistical significance of the comparison, whereas along with known biomarkers, comparing selected new biomarkers in both studies is necessary to show consistency among the databases [56, 88]. Furthermore, the use of only target-positive CLs might restrict the application of methods used earlier such as the waterfall method described in multiple studies [12, 13].

Detection of missense mutation

Another discrepancy that has been pointed out by Haibe-Kains et al. is in the detection of missense mutations in identical CLs. To find the reason behind the discrepancy, Hudson et al. [89] compared missense mutations found in 568 CCLs sequenced by CCLE and the COSMIC, v6 database [23] by the Sanger Institute. Across 1630 mutually sequenced genes, they observed 57.38% conformity (among 45 377 total mutations, 26 038 were in both databases). They also discovered over 400 cold-spots (100bp or larger) in cancer census or kinase genes after analyzing 10 randomly selected CCLE whole-exome sequencing files [89, 90]. These spots were rich in GC nucleotides indicating that the high GC content might result in inadequate sequencing coverage leading to the discrepancy. The other factors that may have affected mutation detection are library preparation, reagents, amplification efficacy, variations in dbSNP filtering, acquisition or loss of mutations and poor reproducibility of data [89]. Using a newly identified PAK4 mutation that lies in GC-rich cold-spot regions, Hudson et al. [89] have found novel driver mutations in known tumor suppressors and oncogenes when specific GC-rich cold-spot regions have been targeted and sequenced. They have argued that discrepancy in pharmacogenomics data is not mainly from mutational profiles, as this mutational status was not crucially incorporated with drug responses [89].

Discontinuous distribution of CCLs

The analysis of two pharmacological profiles (CCLE and GDSC) has brought an important insight, that the distribution of CCL sensitivities is highly discontinuous. This property was obvious considering the fact that a single drug cannot have a similar effect on all the different cancer subtypes, as they are target-specific oncogenic dependent. Besides that, the intrinsic noise of the HTS has caused the drug response variability for inactive compounds and thus has no biological meaning [91]. Therefore, only a handful of CLs can be found, which are drug-sensitive, while the majority of CLs are relatively insensitive to a given drug [76].

Among the 15 drugs common in CCLE and GDSC, 13 of them are dominated by drug-insensitive CLs, making it harder for appropriate pharmacological assessments. In a couple of drugs, the number of CLs that are drug-sensitive in both databases is <5, making any comparison invalid. For most of the drugs, after removing the drug-insensitive CLs, the updated correlation [76] among the drug sensitivities is higher compared with the

correlation values given by [12]. Another change that was mentioned by [56] and later by [76] was that instead of the Spearman correlation coefficient (as used by [12]), they have used the Pearson correlation coefficient because of its higher efficacy in reflecting strong consistent relationship in discontinuous distribution.

Geleher et al. have pointed out that for a given drug, there is little biological variability across the majority of CLs, which has resulted in low correlation between CCLE and GDSC [79]. If the pharmacology of the drugs (e.g. drug nilotinib and BCR-ABL1-targeted CLs) were considered, a valid comparison could have been made. To select the optimal cutoff in the cases where genes of interest were rarely expressed, Safikhani et al. [88] used MCC [92]. Although MCC is a more suitable index for consistency measure than Spearman's or Pearson's correlation coefficient because of its overoptimistic characteristics, low values of MCC for most of the drugs suggest that there are no pertinent interconnections between drug sensitivity variation and MCC estimates [88]. Bouhaddou et al. [93] also found around 85% CLs that are insensitive to the majority of the tested drugs. Characterizing the consistency between data sets with these numbers of inconsistent CLs is difficult.

Adjustment in drug sensitivity values

Maximum drug concentration tested and the range of drug concentrations tested are two metrics that pose a mathematical and analytical challenge in the integration of diverse pharmaceuticals studies [13, 21, 24]. As the number of drug-insensitive CLs is high for the majority of the drugs, extrapolation is required for arriving at drug sensitivity metrics such as half maximal inhibitory concentration (IC_{50}). As the range of tested drug concentrations are different for various databases, the values of drug sensitivities are primarily estimation even when the maximum tested drug concentration is relatively high [12].

The CCLE Consortium and the GDSC Consortium have considered this difference in methodology by capping the IC_{50} value at the maximal drug concentration, but in the process, most of the CLs were capped (for some drugs as high as 98% CLs) and the result was overestimation of the correlation between IC_{50} values [76]. Another disadvantage of capping is elimination of useful drug sensitivity information.

Considering the difference in maximal tested concentrations and the range of tested concentrations, Pozdeyev et al. [91] proposed a new metric, the adjusted AUC. This metric is based on two factors: (i) using sigmoidal curve parameters approximated with a standard logistic regression and (ii) calculating for the range of concentrations only, which is common among the dose-response curves being compared. After the adjustment of the AUC values, CCLE [13] drug sensitivity data have high correlation (0.82) with that of CTRP [21], while GDSC [24] drug sensitivity data have moderate correlation (0.65) with that of CTRP [21]. Bouhaddou et al. [93] also computed a common viability metric (0–100%) across a shared $\log_{10}$ -dose range and calculated the Hill slope and AUC values. Using the new metrics, a better quantitative agreement between CCLE and GDSC has been shown (Pearson correlation coefficient between AUC = 0.61). Safikhani et al. [94] further improved the Hill slope metric with the PharmacGx package [95] to exclude the highly sensitive CLs with flat dose-response curves from being classified as insensitive. With that minor improvement, the sensitivity computation has further improved (Pearson correlation coefficient between AUC = 0.67). But with these common viability metrics,

improvement in consistency is marginal for full drug concentration range [94].

Similar to Pozdeyev et al. [91] and Bouhaddou et al. [93], Mpindi et al. [30] have suggested the drug sensitivity score (DSS) as a drug response metric, defined as a standardized AUC metric, computed from the drug concentration range shared between studies using different curve-fitting algorithms. To show the improvement of this DSS metric, a new data set, FIMM [31], has been compared with the CCLE and GDSC database. After the unification of drug concentration ranges across the CCLE, GDSC and FIMM assays, a markedly higher concordance ($p = 4.2 \times 10^{-5}$, using two-sided Wilcoxon rank-sum test) is observed. Median rank correlation between CCLE and FIMM and GDSC and FIMM drug response data was found to be 0.74 and 0.54, respectively. To validate the result further and find out the reason for this significant improvement (whether the use of same drug dose-response curve modeling or the choice of concentration range), Safikhani et al. used the PharmacGx package [95] and the same curve-fitting algorithm for CCLE, GDSC and FIMM again [96]. Although Safikhani et al. [96] found significantly high correlation between CLs for AUC values computed using a shared CCLE and GDSC concentration range, they have noticed no significantly higher correlation across CLs for AUC values between the CCLE and FIMM. Both Mpindi et al. [30] and Safikhani et al. [96] have agreed that modified AUC values computed on the harmonized concentration range have better correlation between CCLE and GDSC than the published unharmonized AUC values.

Comparison methods

For the purpose of showing inconsistencies between pharmacogenomic databases, Haibe-Kains et al. [12] have reported correlation for different measures between CCLE and GDSC, but inconsistently 'between' CLs for gene expression and 'across' CLs for drug sensitivity. After correcting the anomaly, they have attained a median Spearman's rank correlation coefficient, $r_s = 0.88$ between CLs and $r_s = 0.56$ across CLs for gene expression and median $r_s = 0.62$ between CLs and $r_s = 0.35$ across CLs for AUC [30, 56, 79]. Even after correcting this inconsistency, gene expression data are still significantly more correlated between databases than pharmacological response data [88]. Furthermore, the lower correlation of expression data across CLs raises more doubt about the consistency of the CCLE and GDSC databases. Although the original publications [13, 22] emphasized comparing data both ways, following the same evaluation approach is important for an ideal comparison. In contrast, Safikhani et al. [56] have proposed that the across-CLs comparison is more valid than the between-CLs comparison.

Using a binary sensitivity classification, Bouhaddou et al. [93] tried to evaluate the consistency between CCLE and GDSC databases. For this purpose, all dose-response curves have been curated manually for each study, and then separate support vector machines (SVMs) [97] have been built using a common viability metric. Both SVMs of the two studies have performed well and the two decision boundaries for CCLE and GDSC are similar. Comparison of manual curation between the two studies along with the use of the CCLE SVM to classify GDSC data and vice versa shows high statistically significant consistency (88%) [93]. The inconsistent drug/CL points are within minimal distance away from the decision boundary for 53% of the points in CCLE and 51% of the points in GDSC. These results indicate that the inconsistency between studies is largely because of the information loss because of collapsing a 2D continuous description of drug sensitivity onto a single binary variable. Whereas

Bouhaddou et al. [93] found correlation of 0.69 between the studies, this only occurs for the sensitive drug/CL pairs in either CCLE or GDSC determined by the SVM classifier. Bouhaddou et al. [93] reported a Cohen's kappa (κ) value of 0.53, which is consistent with the MCC value of 0.53 reported in [88]. Safikhani et al. [94] disagreed with Bouhaddou's [93] claim of consistency by referring to the strength of agreement of κ [98]. Moreover, Safikhani et al. [94] claim to have observed no significant improvement because of manual curation over the classification using the common viability metrics when the classifications are stratified by the drugs.

Table 5 summarizes the various studies, which support the consistency of the large pharmacogenomic studies. Furthermore, we have presented in Table 6 a simple instruction protocol that shows the factors of significance that have been taken till now to compare information from different databases.

Effects and remedies for database inconsistencies

These pharmacological databases are being used by numerous research institutes and laboratories to discover molecular mechanism of cancer activity or generate hypothesis for the development of personalized therapy. Inconsistencies in these databases can result in inference of predictive models that have low performance on testing data sets. Standardization of the pharmacological protocol used in the generation of these databases, like assay methods or laboratory conditions, will help the research community. For that purpose, a community-wide consortium effort is necessary. It should be noted that these databases are regularly updated, so a combined effort between groups can reduce the inconsistency. For example, with the existing databases, transfer learning (TL) methodologies [99, 100] could be used when two databases come from two different

domains. In addition, adjusting or changing drug sensitivity metrics [30, 76, 91, 93] or considering biological noise in the published studies can assist in improving consistency.

Common biomarker selection

Most of the recent pharmacogenomic studies use HTSs to collect information from various genomic levels resulting in extremely high-dimensional genomic characterization data sets. Among the genes studied, only a few carry valuable information for drug sensitivity; thus, considering all the genes in the intended learning algorithm could result in model overfitting [7, 19, 101]. The issue of overfitting can be addressed via feature selection, which can be divided into three categories: Filter, Wrapper and Embedded feature selection techniques.

Filter feature selection: One of the most commonly used approaches of feature selection is to use filter methods. The design criteria for filter methods are based on the general statistical characteristics of the data such as statistical independence or correlation measure with the output response. Some common examples of filter methods include (i) ReliefF [7, 101, 102], a computationally inexpensive, robust, noise tolerant technique using k-NN approach for pertinent feature selection but fails to discriminate between redundant features, (ii) minimum redundancy maximum relevance [103], which considers the features with high statistical dependence on output response while minimizing the redundancy in the selected subset, (iii) correlation coefficients between the genomic characteristics and corresponding output responses that emphasize on the individual importance of each feature compared with the response.

Table 5. For Showing consistency of different pharmacological data sets, key factors like aspects or, methods or, sources discussed by different papers have been summarized here

Publication title	Key factors discussed	Database used	Aspect	Method used
Pharmacogenomic agreement between two CCL data sets [76]	(i) Distribution of drug responses (ii) Predictive biomarkers	(i) CCLE (ii) GDSC	(i) AUC (ii) IC ₅₀ (iii) Biomarkers	(i) Pearson correlation (ii) Waterfall analysis (iii) ANOVA (iv) EN
Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies [91]	(i) Drug sensitivity metric	(i) CTRP (ii) CCLE (iii) GDSC	(i) AUC (ii) IC ₅₀ (iii) EC ₅₀	Pearson correlation (ii) Logistic regression
Consistency in large pharmacogenomic studies [79]	(i) Drug targets (ii) Direction of analysis	(i) CCLE (ii) GDSC	(i) Biomarkers (ii) AUC	(i) Spearman correlation (ii) EN
Drug response consistency in CCLE and CGP [93]	(i) Drug sensitivity metrics	(i) CCLE (ii) GDSC	(i) Hill Slope (ii) AUC (iii) IC ₅₀	(i) SVM
Consistency of drug response profiling [30]	(i) Drug sensitivity metric (ii) Experimental protocol (iii) Direction of analysis	(i) CCLE (ii) GDSC (iii) FIMM	(i) AUC	(i) Spearman correlation
Reproducible pharmacogenomic profiling of CCL panels [28]	(i) Identify CLs (ii) Genomic biomarkers (iii) Experimental Setup	(i) CCLE (ii) GDSC (iii) gCSI	(i) AUC (ii) IC ₅₀ (iii) Biomarkers	(i) Two-compound mixture distribution (ii) EN (iii) Univariate regression
Discrepancies in cancer genomic sequencing highlight opportunities for driver mutation discovery [89]	(i) Mutation detection	(i) CCLE (ii) GDSC	(i) Mutation	(i) Whole-exome sequencing
Analysis of variability in high-throughput screening data: applications to melanoma CLs and drug responses [73]	(i) Experimental Setup	TGen (ii) SBP	(i) Dose-response points	(i) ANOVA

Table 6. A simple instruction protocol is presented here that shows the significant factors and procedures to implement those factors for comparing information from different databases

Factors of significance	Procedures
Linear relation [12, 56, 76]	(i) Pearson correlation (ii) Spearman rank correlation (iii) Waterfall analysis (iv) Somers' Dxy rank correlation (v) Mathews correlation coefficient
Biomarker selection [12, 28, 76]	(i) EN (ii) ANOVA
Statistical test [12, 93]	(i) Wilcoxon rank-sum test (ii) Cohen's Kappa (κ) coefficient (iii) SVM classifier
Gene–drug association [12, 56]	(i) Linear regression model (ii) GSEA (iii) Jaccard index

Wrapper feature selection: In contrast to filter feature selection, wrapper techniques incorporate model design in search of the features itself. The selection criteria in wrapper techniques are the predictive performance of the chosen feature subset using a particular model. Here, for a particular feature set $S_p \subset S$, the goodness of fit is evaluated using an appropriate objective function $J(S_p)$, which can be model accuracy measure using correlation coefficient, mean absolute error or mean square error between predicted and actual/experimental responses. Some typical examples of wrapper techniques in drug sensitivity prediction include (i) Sequential Floating Forward Search [104, 105], which considers the selection of an additional feature from the remaining feature set iteratively via cost minimization or reward maximization, while the floating part provides the option for removal of a selected feature if it improves the objective function, (ii) Recursive Feature Elimination [106], which initially fits a model to data with the complete feature set to produce a ranking and recursively

eliminates the lowest ranked features, (iii) Genetic Algorithm Feature Selection (GAFS) [107, 108], an evolution inspired approach where strong selections of features have a higher opportunity to pass their chosen features to offspring via reproduction and weaker selections are eliminated by natural selection.

Embedded feature selection: Often feature selection can be performed as part of the learning process, thus regarded as Embedded feature selection [109, 110]. Compared with the wrapper techniques, embedded approaches incorporate the specific structure of the model to select the relevant features, and therefore, these approaches cannot be separated from the learning process. Frequently, regularization is used for embedded feature selection such as LASSO [111, 112] (penalizes the L_1 norm), Ridge Regression [113, 114] (penalizes the L_2 norm) or EN [115] (penalizes a weighted mix of L_1 and L_2 norms).

As the genomic characterization and drug sensitivity analysis of CCLE and GDSC have followed standard protocols, we would expect that the features selected for these two studies using various feature selection approaches will be similar. In prior studies, the task of biomarker, target or gene selection has been conducted using primarily EN, which is an embedded feature selection approach [76, 79, 93]. In this article, we have explored three representative feature selection approaches to observe which approach provides best idea about the consistency of CCLE and GDSC studies. We have selected one approach from each selection methodologies, ReliefF [102] as Filter Feature selection, GAFS [107] as Wrapper Feature Selection and LASSO [111, 112] as Embedded Feature Selection. These methods were applied separately in CCLE and GDSC for the 15 663 common genes. In the top 100 features selected using these techniques, the number of common features or biomarkers among the data sets is shown in Table 7. For the GAFS approach, the feature set was first reduced using ReliefF to 500 features in both the GDSC and CCLE databases, and we then took the union of these features and applied a Genetic Algorithm to pick the top 100 features.

From Table 7, we observe that there is no general consistency between the numbers of common features selected by

Table 7. The number of common features or biomarkers among the CCLE [13] and GDSC [24] data sets is shown here in the top 100 features selected using ReliefF, genetic algorithm and LASSO

Drug name	Number of common CLs	Correlation coefficient		ReliefF	Genetic algorithm	LASSO
		Pearson	Spearman			
17-AAG	309	0.5411	0.5627	8	14	6
AZD0530/saracatinib	105	0.5756	0.5093	21	15	35
AZD6244/selumetinib	295	0.3982	0.2626	7	10	1
Erlotinib	91	0.4818	0.3922	16	14	38
Lapatinib	98	0.5555	0.4344	25	12	26
Nilotinib	237	0.8872	0.0871	50	12	11
Nutlin-3	310	0.4091	0.3193	9	21	4
Paclitaxel	101	0.3960	0.4192	5	7	49
PD-0325901	306	0.6439	0.5908	16	14	5
PD-0332991	253	0.2507	0.1886	13	9	5
PF2341066/crizotinib	105	0.6190	0.2765	30	10	31
PHA-665752	105	0.0596	−0.1107	5	12	32
PLX4720	304	0.5686	0.2882	33	15	4
Sorafenib	101	0.5205	0.3372	43	11	39
TAE684	105	0.6361	0.4801	25	13	40

Note: Along with that Pearson and Spearman correlation coefficients among the drug sensitivity measure, 'AUC' of CCLE and GDSC are included.

each algorithm. The average number of common features among the top hundred for the three algorithms is 20.4, 12.6 and 21.7 for ReliefF, Genetic Algorithm and LASSO, respectively. We also did not observe any correlation between number of common CLs and the number of common features selected by either ReliefF or Genetic Algorithm-based feature selection algorithms. However, we observed a negative correlation of -0.95 for the number of common features selected by Lasso and the number of common CLs. It appears that with more CLs, the features selected by Lasso for the two databases tend to be disjoint.

8 Data consistency across data sets NCI60, GDSC and CCLE

In this section, we consider the additional data set of NCI60 and compare its consistency with GDSC and CCLE. Figure 1 depicts three Venn diagrams displaying the data intersection between NCI60, GDSC and CCLE data sets. A total of 15 579 genes are common for all three assays as shown in Figure 1A. The NCI60 project receives up to 3000 small molecules per year to be tested, which is reflected in Figure 1B, as it can be seen that the NCI60 project has between 200 and 2000 times more compounds tested, but only 9 of them are common among the three sets. Among the three sets, only 30 CLs are common between the different collections as shown in Figure 1C.

For the common drugs between NCI60 [15], GDSC and CCLE data sets, we have calculated correlation coefficients for drug sensitivity measure IC_{50} . The number of common drugs between CCLE and GDSC, CCLE and NCI60 and NCI60 and GDSC is 16, 10 and 132, respectively. In this article, we have calculated correlation coefficient for three different scenarios.

- Direct correlation:** For a common drug between two data sets, we have considered CLs that have IC_{50} values for both data sets and calculated Pearson or Spearman correlation coefficient using those IC_{50} values.
- Range-adjusted correlation:** In this approach, we have adjusted the range of dose concentration tested, as for two data sets, the doses used for same CL and same drug are often different. When 50% inhibition of a CL is not reached for the tested concentrations, CCLE and NCI60 have reported the maximum concentration tested as the IC_{50} value, whereas GDSC has extrapolated the fitted curve to arrive at the IC_{50} value. As a first step, we have calculated the maximum threshold value by considering the minimum of the maximum concentrations tested for the two data sets for the common drug and converted all IC_{50} values above the maximum threshold value into the maximum threshold value.

Subsequently, we have calculated the minimum threshold value by taking the maximum of the IC_{50} values for the two data sets for the common drug (as minimum doses of NCI60 and GDSC are not available) and converted all IC_{50} values below the minimum threshold value into the minimum threshold value. Finally, we have calculated Pearson or Spearman correlation coefficient using those range-adjusted IC_{50} values.

- Log-converted correlation:** In this approach, we have converted the range-adjusted IC_{50} values by the following equation:

$$\text{Sensitivity} = \frac{(\log_{10}(\text{Maximum threshold}) - \log_{10}(IC_{50}))}{\max((\log_{10}(\text{Maximum threshold}) - \log_{10}(IC_{50})))} \quad (1)$$

The sensitivity values will lie in the range (0, 1). With this conversion, sensitivity of all the insensitive CLs, whose IC_{50} values are equal to the maximum concentration tested, will be 0 and sensitivity of highly sensitive CLs, whose IC_{50} values are very low, will be 1. Subsequently, we have calculated Pearson or Spearman correlation coefficient using these log-converted IC_{50} values.

Figure 2 shows the Pearson (blue) and Spearman (yellow) correlation coefficients for the three different approaches for three different database pairs of CCLE and GDSC, CCLE and NCI60 and NCI60 and GDSC. For the database comparison between NCI60 and GDSC, we have reported a distribution of the correlation coefficients, as the number of common drugs is 132. For the cases where there is no variation among the IC_{50} values of two data sets, correlation coefficient is considered to be 0. It appears that the average correlation for common drugs of CCLE and GDSC has increased after the range adjustment and log conversion. However, the correlation coefficients between NCI60 and CCLE and NCI60 and GDSC are low. The reasons for these low correlations can potentially be because of the limited number of common CLs between NCI60 and the other two databases with most of them being insensitive along with different dose test ranges being used by NCI60 and the other two data sets.

Consistency between databases by considering responses of drug pairs

Existing analysis to show consistency between drug responses of CCLE and GDSC compares the responses of individual drugs that are tested in both studies. These results have been unsatisfactory because of mismatches in experimental protocol or data processing. The call for standardizations of protocols in

Figure 1. Venn diagrams of data intersection between NCI60, CCLE and GDSC data sets.

Figure 2. (A–C) Pearson (blue) and Spearman (yellow) correlation coefficient between common drugs and common CLs of CCLE and GDSC. With range adjustment and normalization, average Pearson correlation of drugs increased considerably (of 16 drugs, 8 has CC > 0.5, which is shown using red reference line of value 0.5). While for NCI60 and CCLE (D–F) and NCI60 and GDSC (G–I), correlation coefficient is generally low. In (G–I), histogram of correlation coefficient is shown as number of common drugs between NCI60 and GDSC is high. In some cases, the correlation coefficient is not considered because of no variation in IC_{50} along with CLs.

previous studies raises the question of whether the dependency structure between drug responses in one study is maintained in another study. Thus, it is important to explore if the dependency structure between drug pairs and cell pairs is maintained between different studies. We consider both linear and nonlinear dependencies for our analysis. The linear dependencies are explored using correlation coefficients and nonlinear dependencies explored using copulas. Note that pairwise consistency explored in this section is being used as a constraint for consistent databases. Consistent databases should maintain pairwise dependency structure, and thus, the analysis of pairwise dependencies can be used a measure of consistency.

Linear dependency between responses of drug pairs

If the only difference between two studies is the drug screening protocol, we can assume that the change in responses from one drug to another will be similar in both the studies, i.e. the dependency structure between the responses of the two drugs will be same. A common way to explore this structural similarity is to check the linear relationship, such as correlation, between the responses for a pair of drugs. For the 15 common drugs between CCLE and GDSC, we have 105 ($= 15C_2$) different drug pairs. Figure 3 shows the Pearson correlation coefficients for drug pairs in CCLE and GDSC jointly. Figure 3 shows that there is a good correlation between the pairwise dependency structures of CCLE and GDSC, i.e. if the responses of a drug pair in CCLE are highly correlated, it is expected that the responses for the same pair in GDSC will also be correlated. The Pearson correlation coefficient of the calculated correlation coefficients between CCLE and GDSC drug pairs is 0.74, indicating that change in responses in same drug pair for both the studies is, for the most part, maintained. In

contrast, correlation between responses of drug pairs is low, which is expected, as two drugs will behave differently if tested on the same CLs.

We have also checked the consistency of drug pairs between the databases through the use of bootstrapping. For each of the 105 drug combinations, we considered a bootstrap sampling of the CCLE AUC values for all CLs that were tested in both studies. We then take the Pearson correlation coefficients of the bootstrap sensitivities and repeat this process 100 times with a new bootstrap sample every iteration. These values are then compared with the correlation coefficient measured in GDSC. If the studies are consistent, the GDSC correlation coefficient is expected to lie within the range of bootstrapped coefficients in the CCLE database. Of the 105 drug combinations, the correlation coefficients of the GDSC samples lie between these ranges for 72 of the cases. The results for all test combinations are shown in Figure 4.

Nonlinear dependency between responses of drug pairs using copulas

Another way to analyze the dependency structure between responses of drug pairs of CCLE and GDSC is by considering nonlinear relationships among the responses [12]. We have used copulas for analyzing this structure, as they can separate the relationship structure in a multivariate probability distribution from the marginal distributions. A copula function [116] is used to represent the dependency structure between multiple random variables without interference from the marginal distributions. Its map of cumulative probability distribution is expressed through the marginal cumulative probability distributions. Let $\Psi_1, \Psi_2, \dots, \Psi_N$ represent N real-valued random variables uniformly distributed on $[0, 1]$. Copula $C: [0, 1]^N \rightarrow [0, 1]$ with parameter θ is stated as:

$$C_\theta(u_1, u_2, \dots, u_N) = P(\Psi_1 \leq u_1, \Psi_2 \leq u_2, \dots, \Psi_N \leq u_N) \quad (2)$$

Figure 3. Pearson correlation coefficients of CCLE and GDSC drug pairs are shown in two dimensions.

Figure 4. For all the 105 possible cases, box plots of Pearson correlation coefficient values for drug pair responses of bootstrapped sets of CCLE are shown here (red plus signs indicate the outliers sets). Along with that corresponding Pearson correlation coefficient values for same drug pair responses of GDSC (green stars) are included to show how many times these correlation lies inside the box.

Sklar's theorem [116] states that the relationship between multivariate cumulative probability distribution $F_X(x_1, x_2, \dots, x_N)$ and marginal cumulative probability distributions $F_i(x_i)$ for $(i \in \{1, 2, \dots, N\})$ is given by:

$$F_X(x_1, x_2, \dots, x_N) = C(F_1(x_1), F_2(x_2), \dots, F_N(x_N)). \quad (3)$$

Copula C is unique [116], whenever the marginal cumulative distributions $(F_i(x))$ are continuous.

Some copulas can be parameterized using only a few parameters; for instance the Clayton copula [117] for a bivariate distribution is defined as follows using parameter ξ :

$$C(u_1, u_2; \xi) = (u_1^{-\xi} + u_2^{-\xi} - 1)^{-1/\xi}; \xi \in (0, \infty). \quad (4)$$

In a similar way, $C(u_1, u_2) = u_1 u_2$ represents the copula that characterizes two independent variables. Some other common forms of parameterized copulas include Gaussian Copula [118], Frank Copula [119], Student's t-copula [120] and Gumbel copula [121]. However, the standard forms of parameterized copulas may not capture all forms of relationships. In that situation, we can consider the use of empirical copulas that are estimated directly from the cumulative multivariate distribution. But one drawback of empirical copula is its high computational complexity

compared with other parameterized copulas, but they can capture a broad range of relationships. We have used Gaussian copulas to represent multivariate dependencies [122] in our analysis.

Analyzing gene expression dependencies using copulas

Using the Spearman rank correlation, it has been shown that the expression profiles for identical CLs of CCLE and GDSC databases are highly correlated. In this section, we illustrate that the dependency captured through copulas between identical CLs of CCLE and GDSC is well maintained. Figures 5 and 6 provide a pictorial representation of the creation of the various copulas. The Frobenius norm between copula of CL i and the copula of CL j for all pairs of CLs is calculated. To estimate whether these generated norms are small or large, we also calculate the Frobenius norm between copula of CL i and the copula of CL j with the order of the genes being randomly permuted as shown in Figure 6. The distribution of these two types of norms of copula differences is shown in Figure 7. Figure 7 clearly indicates that the dependency structure between the two databases is maintained for individual CL gene expressions.

Another way of looking into dependencies is by considering copulas for genes common in both studies. There are 15 663 identical genes between CCLE and GDSC database, which will

Figure 5. Illustration of copula generation with three hypothetical common CLs of CCLE and GDSC with five genes.

Figure 6. Illustration of copula generation with a hypothetical common CL of CCLE and GDSC with five genes that are ordered differently for different cases.

Figure 7. Distribution of Frobenius norm difference of copulas with ordered and disordered genes of identical CLs of CCLE and GDSC database. Mean of Frobenius norm difference of ordered gene case is 0.05, while for disordered gene case, it is 2.12.

gives us 15 663 copulas. The difference between these copulas is compared with difference between copulas of ordered and disordered CLs. The distributions of these two types of differences are shown in Figure 8, where the mean of these two distributions are 0.75 and 1.44, respectively, indicating high nonlinear relationship between genes of CCLE and GDSC database. We note that the Frobenius norm of differences in copulas for ordered CLs is high in Figure 8 as compared with difference in copulas for ordered genes as shown in Figure 7. The behavior is similar to observed before when gene expression correlation coefficient across CLs was found to be lower than gene expression correlation coefficient between CLs.

Analyzing drug response dependencies using copulas

Dependency between drug sensitivities of CCLE and GDSC databases has been investigated using two different methods. First, similar to our studies with gene expression, copulas are generated using responses for 15 common drugs of the two databases

and the Frobenius norm differences calculated. For comparison purposes, we also generate disordered responses of the common drugs to generate disordered copulas that are compared with ordered response copulas. The distribution of Frobenius norm differences, shown in Figure 9, illustrates that the ordered response copulas of 15 common drugs are much more structured or correlated compared with the disordered copulas indicating limited discrepancy among the sensitivities.

In the second approach, we have formulated a copula using the responses for a pair of drugs found in CCLE and compared it with a copula using the responses of the same drug pair in GDSC. To form a comparison distribution, disordered CLs are considered and copulas generated. The process has been repeated for all $15C_2 = 105$ drug pairs. The distribution of differences between ordered and disordered copulas for all 105 cases is shown in Figure 10. Similar to the linear correlation case, we observe that ordered copulas of the drug pairs are much more structured (mean of Frobenius norm differences is 0.23) as

Figure 8. Distribution of Frobenius norm difference of copulas with ordered and disordered CLs of identical genes of CCLE and GDSC database. Mean of Frobenius norm difference of ordered CL case is 0.75, while for disordered CL case, it is 1.44.

Figure 9. Distribution of Frobenius norm differences of copulas with ordered and disordered CLs of identical drugs of CCLE and GDSC database. Mean of Frobenius norm difference of ordered CL case is 0.29, while for disordered CL case, it is 0.76.

Figure 10. Distribution of Frobenius norm differences of copulas of drug pairs with ordered and disordered drug responses of 15 common drugs from CCLE and GDSC database. Mean of Frobenius norm difference of ordered CL case is 0.23, while for disordered CL case, it is 0.53.

compared with disordered copulas (mean of Frobenius norm differences is 0.53). Note that if the comparison distribution would have been created from totally random copulas, the difference between the two distributions would have been even higher. The results reflect that the relationships between a pair of drugs are overall maintained between the two databases.

Conclusion

In this article, we have discussed the different aspects of high-dimensional pharmacogenomics data that have contributed to

variations among databases and examined strategies to analyze the relationships between pharmacogenomics databases. Initially, we have presented a brief overview of different pharmacological databases and drugbanks related to personalized therapy of cancer. We have highlighted the kind of genomic and pharmacological information included in those databases and the processes that have been followed to generate the data.

Although direct comparison among genomic characterizations of CCLE and GDSC has shown consistency, inconsistency is observed for drug sensitivity measures among the two

databases. The experimental protocol and the procedures for curve fitting (subsequently generating responses) have been contributing factors toward the inconsistency. Biological analysis such as identifying genomic predictors and computing normalized enrichment scores has indicated inconsistency among the databases. We applied different feature selection approaches to analyze the number of common features in the top feature sets of CCLE and GDSC and observed limited consistency. However, applying same experimental protocol (like FIMM and CCLE) or using identical drug phenotypes have improved the consistency among databases. In addition to that, by using penalized regression strategy of EN or performing ANOVA test, known genetic biomarkers of sensitivity or resistance have been identified. Furthermore, separating drug-sensitive CLs from drug-insensitive CLs or adjusting drug sensitivity values by considering maximum tested drug concentration and the range of tested drug concentrations has improved consistency significantly. Other than reviewing the existing results in pharmacogenomics database comparisons, we introduced a new approach to explore these databases that have not been considered before. We introduced the concept of copulas to explore nonlinear dependencies between gene expressions or drug responses along with the analysis of maintenance of pairwise dependencies. Copulas were able to capture consistent dependency structures among the gene expression of two databases and between common drugs.

To summarize, we illustrate that pairwise dependencies between drugs are maintained in the two databases of CCLE and GDSC whose consistency analysis has garnered a lot of interest recently. Furthermore, the use of copulas that can capture any form of dependency provided an alternative approach to study the relationships in the two databases. It is expected that with increasing interest in pharmacogenomics, standardization of protocols for pharmacological response measurements will be forthcoming, which in turn will potentially increase the consistency between diverse databases.

Source codes: All the source code to generate the figures and tables are given in <https://github.com/razrahman/Evaluating-the-Consistency.git>. I have also attached some of the preprocessed data, which are required to generate the figures, while link to the couple of big data is provided inside the codes.

Key Points

Direct comparison of drug sensitivity measures of CCLE and GDSC has shown inconsistency, where the experimental protocol and the procedures for curve fitting (subsequently generating responses) have been the primary contributing factors toward the inconsistency.

Separating drug-sensitive CLs from drug-insensitive CLs or adjusting drug sensitivity values by considering maximum tested drug concentration and the range of tested drug concentrations has improved consistency significantly.

Nonlinear correlation measure copula was able to capture consistent dependency structures among the gene expression of two databases and between common drugs.

Funding

This work was supported by National Institutes of Health (NIH) (grant number R01GM122084).

References

- Altman RB, Flockhart D, Goldstein DB. *Principles of Pharmacogenetics and Pharmacogenomics*. Cambridge: Cambridge University Press, 2012.
- Adams MD, Kelley JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 1991;**252**(5013):1651–6.
- Sinsheimer RL. The Santa Cruz workshop-may 1985. *Genomics* 1989;**5**(4):954–6.
- Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med* 2010;**363**(4):301–4.
- Kannel WB, McGee DL. Diabetes and cardiovascular disease: the framingham study. *JAMA* 1979;**241**(19):2035–8.
- Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med* 2011;**17**(3):297–303.
- Pal R. *Predictive Modeling of Drug Sensitivity*. London, UK: Academic Press, 2016.
- Sharma SV, Haber DA, Settleman J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer* 2010;**10**(4):241–53.
- Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014;**32**(12):1202–12.
- Rahman R, Haider S, Ghosh S, Pal R. Design of probabilistic random forests with applications to anticancer drug sensitivity prediction. *Cancer Inform* 2016;**14**(Suppl 5):57–73.
- Rahman R, Matlock K, Ghosh S, Pal R. Heterogeneity aware random forest for drug sensitivity prediction. *Sci Rep* 2017;**7**(1):11347.
- Haibe-Kains B, El-Hachem N, Birkbak NJ, et al. Inconsistency in large pharmacogenomic studies. *Nature* 2013;**504**(7480):389–93.
- Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**(7391):603–7.
- Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;**483**(7391):570–5.
- Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000;**24**(3):227–35.
- Marioni JC, Mason CE, Mane SM, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**(9):1509–17.
- Eckel-Passow JE, Atkinson EJ, Maharjan S, et al. Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. *BMC Bioinform* 2011;**12**(1):220.
- Rahman R, Pal R. Analyzing drug sensitivity prediction based on dose response curve characteristics. In: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, Las Vegas, 2016, 140–143.
- De Niz C, Rahman R, Zhao X, Pal R. Algorithms for drug sensitivity prediction. *Algorithms* 2016;**9**(4):77.
- Basu A, Bodycombe NE, Cheah JH, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013;**154**(5):1151–61.
- Seashore-Ludlow B, Rees MG, Cheah JH, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov* 2015;**5**(11):1210–23.
- Yang W, Soares J, Greninger P, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker

- discovery in cancer cells. *Nucleic Acids Res* 2012;**41**(D1):D955–61.
23. Forbes SA, Bindal N, Bamford S, et al. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2011;**39**(Database):D945–50.
 24. Iorio F, Knijnenburg TA, Vis DJ, et al. A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**(3):740–54.
 25. Daemen A, Griffith OL, Heiser LM, et al. Modeling precision treatment of breast cancer. *Genome Biol* 2013;**14**(10):R110.
 26. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**(7216):1061–8.
 27. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas pan-cancer analysis project. *Nat Genet* 2013;**45**(10):1113–20.
 28. Haverty PM, Lin E, Tan J, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* 2016;**533**(7603):333–7.
 29. Klijn C, Durinck S, Stawiski EW, et al. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol* 2015;**33**(3):306–12.
 30. Mpindi JP, Yadav B, Östling P, et al. Consistency in drug response profiling. *Nature* 2016;**540**(7631):E5–6.
 31. Pemovska T, Kontro M, Yadav B, et al. Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discov* 2013;**3**(12):1416–29.
 32. Hook KE, Garza SJ, Lira ME, et al. An integrated genomic approach to identify predictive biomarkers of response to the aurora kinase inhibitor pf-03814735. *Mol Cancer Ther* 2012;**11**(3):710–19.
 33. Fallahi-Sichani M, Moerke NJ, Niepel M, et al. Systematic analysis of BRAF v 600e melanomas reveals a role for JNK/C-JUN pathway in adaptive resistance to drug-induced apoptosis. *Mol Syst Biol* 2015;**11**(3):797.
 34. Koleti A, Terryn R, Stathias V, et al. Data portal for the Library of Integrated Network-Based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res* 2018;**46**(D1):D558–66.
 35. International Cancer Genome Consortium, Hudson TJ, Anderson W, et al. International network of cancer genome projects. *Nature* 2010;**464**(7291):993–8.
 36. Zhang J, Baran J, Cros A, et al. International Cancer Genome Consortium data portal—a one-stop shop for cancer genomics data. *Database* 2011;**2011**(0):bar026.
 37. Wishart DS, Feunang YD, Guo AC, et al. Drugbank 5.0: a major update to the Drugbank database for 2018. *Nucleic Acids Res* 2018;**46**(D1):D1074–82.
 38. Siramshetty VB, Eckert OA, Gohlke B-O, et al. Superdrug2: a one stop resource for approved/marketed drugs. *Nucleic Acids Res* 2018;**46**(D1):D1137–43.
 39. Goede A, Dunkel M, Mester N, et al. Superdrug: a conformational drug database. *Bioinformatics* 2005;**21**(9):1751–3.
 40. Cotto KC, Wagner AH, Feng YY, et al. Dgidb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res* 2018;**46**:D1068–D1073.
 41. Russ AP, Lampel S. The druggable genome: an update. *Drug Discov Today* 2005;**10**(23–24):1607–10.
 42. Liu Y, Wei Q, Yu G, et al. DCDB 2.0: a major update of the drug combination database. *Database* 2014;**2014**:bau124.
 43. Whirl-Carrillo M, McDonagh EM, Hebert JM, et al. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012;**92**(4):414–17.
 44. Ursu O, Holmes J, Knockel J, et al. Drugcentral: online drug compendium. *Nucleic Acids Res* 2017;**45**(D1):D932–9.
 45. Forbes SA, Beare D, Boutselakis H, et al. Cosmic: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;**45**(D1):D777–83.
 46. Szklarczyk D, Morris JH, Cook H, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res* 2017;**45**:D362–D368.
 47. Backman TWH, Cao Y, Girke T. ChEMBL tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Res* 2011;**39**:W486–91.
 48. Keenan AB, Jenkins SL, Jagodnik KM, et al. The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations. *Cell Syst* 2018;**6**:13–24.
 49. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: 1 000 platform and the first 1, 000, 000 profiles. *Cell* 2017;**171**(6):1437–52.
 50. Napolitano F, Sirci F, Carrella D, di Bernardo D. Drug-set enrichment analysis: a novel tool to investigate drug mode of action. *Bioinformatics* 2016;**32**(2):235–41.
 51. Brown PO, Botstein D. Exploring the new world of the genome with dna microarrays. *Nat Genet* 1999;**21**(Suppl 1):33–7.
 52. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet* 2012;**13**(7):505–16.
 53. Crawford EL, Weaver DA, Willey JC. Development of a standardized, quantitative microarray for gene expression measurement. *Proc Amer Assoc Cancer Res* 2004;**64**(Suppl 7):379.
 54. Zhou YH, Raj VR, Siegel E, Yu L. Standardization of gene expression quantification by absolute real-time qRT-PCR system using a single standard for marker and reference genes. *Biomark Insights* 2010;**5**:79–85.
 55. Weis BK. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2005;**2**(5):351–6.
 56. Safikhani Z, Smirnov P, Freeman M, et al. Revisiting inconsistency in large pharmacogenomic studies. *F1000Res* 2016;**5**:2333.
 57. Safikhani Z, El-Hachem N, Quevedo R, et al. Assessment of pharmacogenomic agreement. *F1000Res* 2016;**5**:825.
 58. Papillon-Cavanagh S, De Jay N, Hachem N, et al. Comparison and validation of genomic predictors for anticancer drug sensitivity. *J Am Med Inform Assoc* 2013;**20**(4):597–602.
 59. Jang IS, Neto EC, Guinney J, et al. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac Symp Biocomput* 2014:63–74.
 60. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;**85**(3):257–68.
 61. Hatzis C, Bedard PL, Birkbak NJ, et al. Enhancing reproducibility in cancer drug screening: how do we move forward? *Cancer Res* 2014;**74**(15):4016–23.
 62. Harris MA, Clark J, Ireland A, et al. The gene ontology (go) database and informatics resource. *Nucleic Acids Res* 2004;**32**:D258–61.
 63. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**(1):25–9.
 64. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**(43):15545–50.

65. Rao JS, Liu H. Discordancy partitioning for validating potentially inconsistent pharmacogenomic studies. *Sci Rep* 2017; 7(1):15169.
66. Weinstein JN, Lorenzi PL. Cancer: discrepancies in drug sensitivity. *Nature* 2013;504(7480):381–3.
67. Wright Muelas M, Ortega F, Breitling R, et al. Rational cell culture optimization enhances experimental reproducibility in cancer cells. *Sci Rep* 2018;8:3029.
68. Celltiter Promega. 96[®] aqueous one solution cell proliferation assay. *Technical Bulletin*. Madison, WI: Promega, 2005.
69. Hannah R, Beck M, Moravec R, Riss T. Celltiter-glo luminescent cell viability assay: a sensitive and rapid method for determining cell viability. *Cell Notes* 2001;2:11–13.
70. Greshock J, Bachman KE, Degenhardt YY, et al. Molecular targ32et class is predictive of in vitro response profile. *Cancer Res* 2010;70(9):3677–86.
71. Chan GKY, Kleinheinz TL, Peterson D, Moffat JG. A simple high-content cell cycle assay reveals frequent discrepancies between cell number and ATP and MTS proliferation assays. *PLoS One* 2013;8(5):e63583.
72. Gilbert DF, Boutros M. A protocol for a high-throughput multiplex cell viability assay. *Methods Mol Biol* 2016;1470: 75–84.
73. Ding KF, Finlay D, Yin H, et al. Analysis of variability in high throughput screening data: applications to melanoma cell lines and drug responses. *Oncotarget* 2017;8(17):27786–99.
74. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33(1):1–22.
75. Ein-Dor L, Kela I, Getz G, et al. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 2005;21(2): 171–8.
76. Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* 2015;528(7580):84–7.
77. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970;12(1):55–67.
78. St L, Wold S. Analysis of variance (ANOVA). *Chemometr Intell Lab Syst* 1989;6(4):259–72.
79. Geleher P, Gamazon ER, Seoighe C, et al. Consistency in large pharmacogenomic studies. *Nature* 2016;540(7631):E1–2.
80. Rix U, Hantschel O, Dürnberger G, et al. Chemical proteomic profiles of the BCR-ABL inhibitors imatinib, nilotinib, and dasatinib reveal novel kinase and nonkinase targets. *Blood* 2007;110(12):4055–63.
81. Konecny GE, Pegram MD, Venkatesan N, et al. Activity of the dual kinase inhibitor lapatinib (gw572016) against her-2-overexpressing and trastuzumab-treated breast cancer cells. *Cancer Res* 2006;66(3):1630–9.
82. Kelland LR, Sharp SY, Rogers PM, et al. Dt-diaphorase expression and tumor cell sensitivity to 17-allyl amino, 17-demethoxygeldanamycin, an inhibitor of heat shock protein 90. *J Natl Cancer Inst* 1999;91(22):1940–9.
83. Solit DB, Garraway LA, Pratilas CA, et al. Braf mutation predicts sensitivity to MEK inhibition. *Nature* 2006;439(7074): 358–362.
84. Dry JR, Pavey S, Pratilas CA, et al. Transcriptional pathway signatures predict mek addiction and response to selumetinib (azd6244). *Cancer Res* 2010;70(6):2264–73.
85. Tsai J, Lee JT, Wang W, et al. Discovery of a selective inhibitor of oncogenic B-RAF kinase with potent antimelanoma activity. *Proc Natl Acad Sci USA* 2008;105(8):3041–6.
86. Müller CR, Paulsen EB, Noordhuis P, et al. Potential for treatment of liposarcomas with the mdm2 antagonist nutlin-3a. *Int J Cancer* 2007;121(1):199–205.
87. Timm A, Kolesar JM. Crizotinib for the treatment of non-small-cell lung cancer. *Am J Health Syst Pharm* 2013;70(11): 943–7.
88. Safikhani Z, El-Hachem N, Smirnov P, et al. Safikhani et al. reply. *Nature* 2016;540(7631):E2–4.
89. Hudson AM, Yates T, Li Y, et al. Discrepancies in cancer genomic sequencing highlight opportunities for driver mutation discovery. *Cancer Res* 2014;74(22):6390–6.
90. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14(2): 178–92.
91. Pozdeyev N, Yoo M, Mackie R, et al. Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. *Oncotarget* 2016;7(32):51619.
92. Matthews BW. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta Protein Struct* 1975;405(2):442–51.
93. Bouhaddou M, DiStefano MS, Riesel EA, et al. Drug response consistency in CCLE and CGP. *Nature* 2016;540(7631):E9–10.
94. Safikhani Z, El-Hachem N, Smirnov P, et al. Safikhani et al. reply. *Nature* 2016;540(7631):E11–12.
95. Smirnov P, Safikhani Z, El-Hachem N, et al. Pharmacogx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* 2016;32(8):1244–6.
96. Safikhani Z, El-Hachem N, Smirnov P, et al. Safikhani et al. reply. *Nature* 2016;540(7631):E6–8.
97. Cortes C, Vapnik V. Support vector networks. *Mach Learn* 1995;20(3):273–97.
98. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159–74.
99. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010;22(10):1345–59.
100. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data* 2016;3(1):9.
101. Rahman R, Otridge J, Pal R. Integratedmrf: random forest-based framework for integrating prediction from different data types. *Bioinformatics* 2017;33(9):1407–10.
102. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of Relief and Rrelief. *Mach Learn* 2003;53(1/2): 23–69.
103. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 2005;27(8):1226–38.
104. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. *Pattern Recognit Lett* 1994;15(11):1119–25.
105. Berlow N, Davis LE, Cantor EL, et al. A new approach for prediction of tumor sensitivity to targeted drugs based on functional data. *BMC Bioinformatics* 2013;14(1):239.
106. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;23(19): 2507–17.
107. Chaikla N, Qi Y. Genetic algorithms in feature selection. In 1999 IEEE International Conference on Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings, Tokyo, Japan, Vol. 5. 1999, 538–40. IEEE.
108. Soufan O, Klefogiannis D, Kalnis P, Bajic VB. Dwfs: a wrapper feature selection tool based on a parallel genetic algorithm. *PLoS One* 2015;10(2):e0117988.

109. Alshahrani M, Soufan O, Magana-Mora A, Bajic VB. Dannp: an efficient artificial neural network pruning tool. *PeerJ Comput Sci* 2017;**3**:e137.
110. Mayer J, Rahman R, Ghosh S, Pal R. Sequential feature selection and inference using multi-variate random forests. *Bioinformatics* 2018;**34**:1336–44.
111. Robert T. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Methodol* 1996;**34**:267–88.
112. Park H, Imoto S, Miyano S. Recursive random lasso (Rrlasso) for identifying anti-cancer drug targets. *PLoS One* 2015;**10**(11):e0141869.
113. Tikhonov AN. Solution of incorrectly formulated problems and the regularization method. *Sov Meth Dokl* 1963;**4**: 1035–8.
114. Neto EC, Jang IS, Friend SH, Margolin AA. The stream algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications to pharmacogenomic prediction of cancer cell line sensitivity. *Pac Symp Biocomput* 2014:27–38.
115. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005;**67**(2):301–20.
116. Sklar M. *Fonctions de répartition à n dimensions et leurs marges*. Paris: Université Paris 8, 1959.
117. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Int Stat Rev* 1978;**65**(1):141–51.
118. Lee L. Generalized econometric models with selectivity. *Econometrica* 1983;**51**(2):507–12.
119. Frank MJ. On the simultaneous associativity of $f(x, y)$ and $x+y - f(x, y)$. *Aeq Math* 1979;**19**(1):194–226.
120. Demarta S, McNeil AJ. The t copula and related copulas. *Int Stat Rev* 2007;**73**(1):111–29.
121. Gumbel EJ. Distributions des valeurs extremes en plusieurs dimensions. *Publ Inst Statist Univ Paris* 1960;**9**:171–3.
122. Haider S, Rahman R, Ghosh S, Pal R. A copula based approach for design of multivariate random forests for drug sensitivity prediction. *PLoS One* 2015;**10**(12):e0144490.