ELSEVIER

feature

A guide for bioinformaticians: ‘omics-based drug discovery for precision oncology

Jihyeob Mun[1], **Gildon Choi**[2], gchoi@krict.re.kr **and Byungho Lim**[2], lbh82@krict.re.kr

[1] Center for Supercomputing Applications, Division of National Supercomputing R&D, Korea Institute of Science and Technology Information (KISTI), Daejeon, Republic of Korea

[2] Research Center for Drug Discovery Technology, Division of Drug Discovery Research, Korea Research Institute of Chemical Technology, Daejeon, Republic of Korea

Bioinformatics-centric drug development is inevitable in the era of precision medicine. Clinical ‘omics information, including genomics, epigenomics, transcriptomics, and proteomics, provides the most comprehensive molecular landscape in which each patient’s pathological history is delineated. Hence, the capability of bioinformaticians to manage integrative ‘omics data is crucial to current drug development. Bioinformatics can accelerate drug development from initial time-consuming discoveries to the clinical stage by providing information-guided solutions. However, many bioinformaticians do not have opportunities to participate in drug discovery programs. As a starting point for bioinformaticians with no prior drug development experience, here we discuss bioinformatics applications during drug development with a focus on working-level omics-based methodologies.

Introduction

The legacy of past genomics efforts is the establishment of the human reference genome and the identification of genomic variability and disease associations. The findings have led to the widespread acceptance of the concept of precision medicine, but the lack of fast, cost-effective technologies to delineate whole genomes has delayed the clinical implementation of precision medicine. The latest advance in next-generation sequencing (NGS) enabled the sequencing of >50 human genomes per run at a cost of < S$1000 per genome. Thus, a challenge to realizing precision medicine is being resolved, improving the prospect of precision oncology.

Nonetheless, other challenges remain. Although NGS-based international projects [The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC)] have considerably increased the number of actionable cancer mutations [1], the functional impact of mutations on drug targets is largely unknown. Recent base editor technologies allow high-throughput functional characterization of mutations by systematically substituting C•G→T•A and A•T→G•C, which account for more than half of pathogenic mutations [2], but technical limitations remain. Moreover, the low efficiency of traditional drug development is inadequate to address the myriad therapeutic targets. Given that the availability of new drugs tailored to individuals is crucial to support precision medicine, this low efficiency could be an obstacle. The low efficiency is attributed to the intrinsic difficulties of drug development itself, including clinically relevant target-indication selection, efficacy and safety issues with compounds, drug resistance, and the absence of biomarkers. To address these issues,

Corresponding authors: Choi, G. (gchoi@krict.re.kr), Lim, B. (lbh82@krict.re.kr)

https://doi.org/10.1016/j.drudis.2020.08.004

Please cite this article in press as: Mun, J. et al. A guide for bioinformaticians: ‘omics-based drug discovery for precision oncology, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.08.004

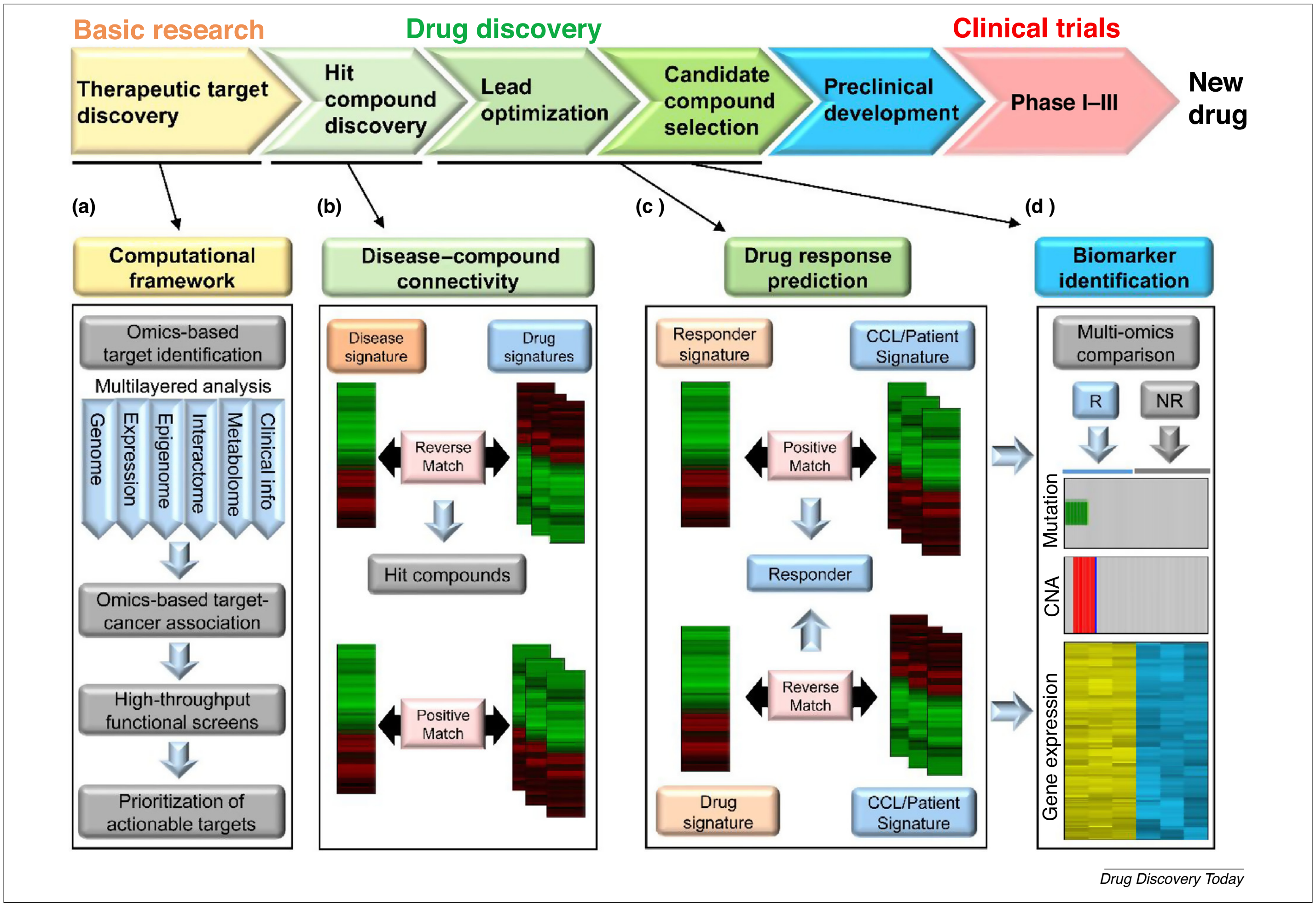


FIGURE 1

'Omics-based approaches in conventional drug development. **(a)** A computational framework for target selection that interrogates multilayered data. **(b)** Transcriptome-based hit discovery that interrogates the transcriptional relationship between diseases and drugs. 'Reverse Match' ('Positive Match'): negative (positive) correlations between disease signatures and drug signatures. **(c)** Transcriptome-based drug sensitivity prediction by assessing the connectivity between responder signatures or drug signatures and CCL/patient signatures. **(d)** Omics-based biomarker identification. Biomarker candidates, including mutations, copy-number alterations (CNAs), and/or gene expression alterations, can be identified through comparative multiomics analyses between responders (R) and nonresponders (NR) derived from sensitivity prediction described in (c). Abbreviation: CCL, cancer cell line.

collaborative efforts in multidisciplinary fields are fundamental. However, we believe that bioinformatics offers rational solutions to some of the issues, as evidenced by research indicating that genetically supported targets can double the success rate of clinical development [3], and that genomics has the potential to improve the odds of developmental success [4].

Target selection

The major impact of bioinformatics on drug development lies in the identification of therapeutic targets for any disease, provided that sufficient data are available. The importance of precise targets cannot be overemphasized because clinically irrelevant targets are the main cause of failure resulting from a lack of efficacy in Phase II clinical trials [5]. Given that targets are sometimes biomarkers, correct target selection is a prerequisite for successful clinical trials.

In conventional drug development, targets are selected in several ways: (i) literature studies; (ii) characterization of individual genes; and (iii) fast-following already-approved targets or those under development. These approaches are still useful but depend on nonsystematic, gene-by-gene methodologies. Challenges can arise with these approaches if doubt exists regarding the justification of target prioritization; that is, why one target is preferentially selected before others. In the case of the fast-following approach, a concern is that considerable development efforts are limited to a few promising targets. By contrast, 'omics-based approaches facilitate systematic target selection by achieving rational target prioritization and ensuring clinical relevance if proper computational frameworks are organized based on the following analytical features (Fig. 1a): (i) multilayered data interrogation; (ii) interpretation by high-throughput functional screening data; and (iii) target prioritization.

Analytical strategies for target selection

Target-disease associations can be interrogated with the following omics-based analytical points (Table 1).

Genome

Persistent genomic alterations have a significant impact on cancer pathogenesis. The main goal is to isolate cancer-associated driver variants by removing nonpathogenic passengers [6]: (i) recurrent mutations: as evidence of selective

Please cite this article in press as: Mun, J. et al. A guide for bioinformaticians: 'omics-based drug discovery for precision oncology, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.08.004

TABLE 1

Omics-based analytical points for target selection.

Data type/source	Analytical points
Genome: TCGA, ICGC, NCBI (GEO, SRI)	Identification of driver events: recurrent mutation, evolutionary action, functional prediction, prognosis (wild-type versus mutant)
Transcriptome: LINCS, GTEx, TCGA, ICGC, NCBI (GEO, SRI)	Gene/gene-set expression changes, differential expression in clinical subtypes, disease(tissue)-specific expression, gene-set/clustering/network-based analysis, prognosis (high versus low)
Epigenome: Epigenome roadmap, ENCODE, TCGA, ICGC, NCBI (GEO, SRI)	Signal changes: association with regulatory elements, DNA methylation and histone modification: signal changes and clustering analysis, (super)enhancer (cancer-type specificity), integration with transcriptomes
Proteome: CPTAC, ICPC, PRIDE	Protein/post-translation level changes, differential expression in clinical subtypes, disease(tissue) specificity, integration with genomes/transcriptomes, gene-set/clustering/network-based analysis
Metabolome: HMDB, MetaboLights	Level changes, differential concentrations in clinical subtypes, oncometabolites for targets, indications, and biomarkers
Interactome: BioPlex, BioGRID, NURSA, OMNIPATH	Core nodes in interactome maps, protein–protein interactions, transcription factor–DNA interactions
Phenome: DepMap, CTRP, GDSC	Genome-wide functional characterization, Integration of perturbagen efficacy, phenotype, and ‘omics data

pressure, two patterns of recurrent mutations are observed: gene-level recurrent mutations indicate significant mutational prevalence over the background mutation rate in the same gene, which can be detected by tools, such as MutSigCV [7]; and mutation hotspots, locations of repeated mutation of the same amino acid residue, tend to elicit oncogenic gain-of-function activity and are targetable by small-molecule inhibitors [8]; (ii) mutational expansion: cells harboring driver mutations that confer a growth advantage expand among heterogeneous intratumoral clones [9]. This evolutionary action is determined by increasing variant allele frequency during cancer progression [10]. The best method to interrogate this evolution involves the single-cell sequencing of temporal and/or spatial series of samples, which provides high-resolution mutational dynamics [11]; (iii) functional prediction: the biophysical consequences of variants on target proteins can be predicted by several algorithms (SIFT, PolyPhen2, Condel, FATHMM, and MutationTaster) using the Variant Effect Predictor [12]; and (iv) prognosis: the clinical relevance of mutated targets is assessed by comparing the survival rates of patients harboring the wild-type or mutant gene.

Transcriptome

The transcriptome is the most universal and largest data source because of advanced technologies and its molecular role as the intermediary between DNA and protein. Genomic information alone is insufficient because of the frequent lack of actionable mutations or the undruggable nature of key mutated proteins (e.g., RAS and MYC) [13]. A study showing that cancer dependency is best predicted by transcriptomes increased the value of transcriptome-dependent target selection [14]: (i) differentially expressed genes (DEGs): DEGs in clinical subtypes (e.g., normal versus cancer) are fundamental for identifying nonmutated oncogenic addiction [15]. DEGs are the input for gene-set analyses [Enrichr [16] and Gene Set Enrichment Analysis (GSEA) [17]] evaluating the key pathways/mechanisms of targets; (ii) gene-set analysis: because most genes exhibit correlated/clustered transcriptional alterations, specifying driver changes based on DEGs only is difficult. Thus, gene-set analysis is more valuable. Unsupervised clustering of cohort data can simultaneously classify transcriptomes and patients, facilitating the functional annotation of coexpressed targets and patient stratification for treatment [18]. The clinical relevance of coexpressed targets is assessed by survival analyses of classified patients. A gene-coexpression network can be constructed to detect high-priority targets that act as master regulators and key pathways [19].

Proteome

Proteins as the major target molecule of drugs are important. Reverse-phase protein array from the TCGA covering the main druggable cancer pathways facilitates an assessment of whether targets are activated. Proteome data are rapidly being acquired using recently developed mass spectrometry (MS). The Clinical Proteomic Tumor Analysis Consortium (CPTAC) data generated using TCGA cohorts highlight integrative analyses, such as proteogenomics, promoting new therapeutic opportunities [20]. Moreover, phosphoproteomes and glycoproteomes from CPTAC and Proteomics Identifications Database (PRIDE) help to identify activated targets and treatable patients [21].

Epigenome

Few clinical uses of epigenomic data exist, but epigenome-driven patient classification is valuable for stratified medicine [22]: (i) enhancer associations: because enhancers define tissue specificity, ChIP-seq analysis of enhancer marks (H3K27ac and H3K4me1) can link enhancer-associated targets to particular tissue lineages [23]. Super-enhancers that define cancer master regulators can facilitate the identification of high-priority targets [24], especially in patients without actionable mutations. The combined information about tissue-specific enhancers (e.g., Roadmap Epigenomics Project) and their interacting genes (e.g., Hi-C data) can match targets with indications [25]; (ii) differentially methylated CpGs (DMCs): DMCs are examined as causal mechanisms of DEGs. Unsupervised clustering of DMCs classifies potential targets and patient subtypes [26]. For example, hypermethylated subtypes in gastric cancer showed higher immunoreactivity, facilitating target discovery and patient selection for immunotherapy [27].

Metabolome

Metabolite profiles are functional readouts with strong phenotypic correlations, thus serving as noninvasive biomarkers for precision medicine [28]. Identification of patients with addiction of a given metabolite can simultaneously identify targets associated with a specific enzymatic reaction and the metabolite as a biomarker (e.g., oncometabolite D-2-hydroxyglutarate resulting from IDH1/2 mutations in glioma [29]).

Interactome

Biological functions are orchestrated by biophysical interactions among DNA, RNA, and

Please cite this article in press as: Mun, J. et al. A guide for bioinformaticians: ‘omics-based drug discovery for precision oncology, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.08.004

proteins [30]. Thus, drug discovery research has focused on small molecules that perturb interactions (e.g., multimeric proteins, and enzyme–substrate and receptor–ligand pairs); (i) protein–protein interactions: recent yeast two-hybrid systems and high-throughput affinity-purification followed by MS have systematically generated human interactome maps (e.g., BioPlex) [30]. Targets can be selected based on the topological properties of essential nodes in interactome networks [31]; (ii) transcription factor (TF)–DNA interaction: combined ChIP-seq data reveal co-occupancy of multiple TFs on regulatory elements [32], revealing TF interactomes as a targeting strategy. Although targeting TFs is challenging, the recent proteolysis targeting chimera (PROTAC) technology induces the degradation of undruggable proteins, including TFs, by hijacking the endogenous E3 ligase and ubiquitin proteasome system [33].

After the 'omics-based interrogation described earlier, another key challenge is to prioritize target–disease associations by defining robust scoring systems. Previously, the integrative platform OpenTargets adopted a scoring scheme that aggregates interrogated data through a four-tier process, including evidence scores, data source scores, data type scores, and overall scores [34]. An association score per each evidence was calculated by considering the confidence and strength of target–disease associations, and aggregation of the resulting association scores was implemented by the sum of the harmonic progression of each score [35]. Another ranking scheme from Pharos estimates the sum of the cumulative probabilities across all data sources from Harmonizome [36], which is defined as the Data Availability Score [37].

High-throughput functional characterization

Prioritized targets should be supported by functional assessments because a high rank does not guarantee a causal relationship between cancer and the target. The most powerful strategy is the genome-wide loss-of-function screen using short hairpin (sh)RNA or short guide (sg)RNA libraries to systematically examine cancer dependency profiles. The DepMap data constitute the Broad Institute Project Achilles (~94 000 shRNAs targeting ~17 000 genes across 501 cell lines) [14] and the Novartis DRIVE (~158 000 shRNAs targeting ~8000 genes across 397 cell lines) [38]. To analyze these data, the impact of targets on cancer cell survival is assessed by barcode sequencing. The effect of cell insertion of each shRNA/sgRNA with a unique barcode is determined by quantification of depleted barcodes relative to the original pooled library. Significant depletion indicates functional implications of targets in cell viability. These data can be analyzed and applied as follows: (i) correction: substantial off-target effects of shRNAs (mostly miRNA-seed effects) can be corrected using computational tools. DEMETER isolates the on-target effects of shRNAs by removing inferred miRNA-seed effects [14]. The current version, DEMETER2, outperformed other tools in estimating absolute and relative target dependencies [39]; (ii) overcoming undruggability: ~80% of targets with strong cancer dependency were estimated to be undruggable [14]. Undruggability is overcome by targeting other proteins engaged in the same pathways/complexes. Correlated dependency profiles in DepMap combined with interactome maps can select targets that bypass undruggability [14]. Another approach to overcome undruggability is performing genome-wide synthetic-lethal screens (e.g., the synthetic lethality of SAE2 with undruggable Myc) [40]; and (iii) biomarker–target pairs: DepMap integrated with 'omics information of cell lines (CCLE [41]) is useful for selecting biomarker-matched functional targets.

Hit discovery

Following target selection, compound screening assays are performed to discover drug-like chemical structures (termed 'hit compounds') with the desired activity or binding affinity to the target. A representative assay is in vitro high-throughput screening (HTS) using chemical libraries to identify hits without prior knowledge of chemical classes. Although these assays are automated, they can be laborious because active recombinant proteins and high-quality assays are needed for these repetitive experiments with 10^6–10^7 compounds [42].

By contrast, computational approaches save time and money, compensating for the weakness of experimental assays. Since the computer-aided development of viracept in 1997, cheminformatics has served as an indispensable method in drug development [43]. Virtual screening (VS) is a major technique used to identify drug candidates among *in silico* libraries by two types of methodology: (i) structure based: rational drug design is achieved using known 3D structure of target proteins (e.g., molecular docking); and (ii) ligand based: active compounds are used as query templates to identify new chemical entities with similar properties [43]. These approaches have improved hit discovery rates (>10% by VS versus 0.01–0.14% by HTS) [44]. A recent study successfully performed structure-based ultralarge library docking of 170 million virtual compounds, which overwhelmed HTS [45]. Details of cheminformatics-based drug discovery can be found elsewhere [43,46].

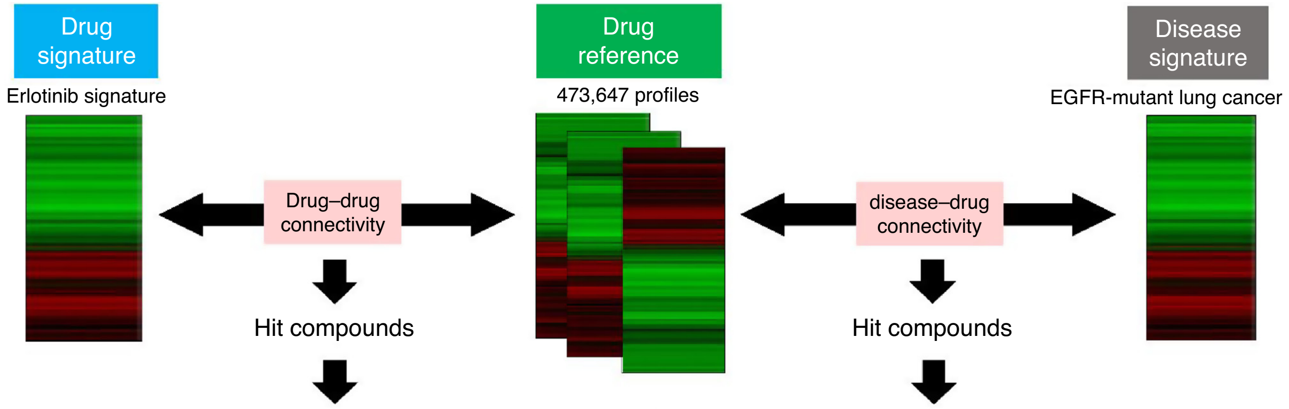
Bioinformatics-based discoveries are based on biological/molecular phenotypic activity. The representative strategy is transcriptome-based approaches adopting the basic concept of the Connectivity Map [47], where similarities between genes, diseases, and drugs are examined through pattern matching of transcriptional profiles. In contrast to conventional target-first discovery, transcriptome-based approaches utilize gene expression signatures as inputs reflecting the molecular/phenotypic activity of diseases and drugs (Fig. 1b) [48]: (i) disease gene expression signatures: molecular symptoms of cancer can be represented by differential gene expression between normal and cancer (termed the 'disease signature'). Various disease signatures can be created based on clinical subtypes or individual patients of interest; (ii) drug-induced gene expression signatures: drug treatment can induce differential gene expression relative to vehicle (termed the 'drug signature'). Library of Integrated Network-based Cellular Signatures (LINCS) has established >1.3 million transcriptome profiles encompassing several perturbagens (19 811 compounds, 18 493 shRNAs, 3462 cDNAs, and 314 biologics) across 3–77 cell lines [49]. The data were generated using a high-throughput bead-array L1000, which measures the expression levels of only 978 landmark genes to infer 82% of the whole transcriptome [49]. This cost-effective method enables the establishment of customized unique drug signatures with extensive coverage to boost novel discoveries.

To retrieve hit compounds, the connectivity of these two signatures is interrogated using nonparametric rank-ordered Kolmogorov–Smirnov statistics [50] (Fig. 1b); 'reverse match' denotes the identification of potential hits that can reverse a disease state to the normal state, whereas 'positive match' indicates exaggeration of a disease state by drug candidates. As a proof of principle, a study demonstrated that the reversal potency of disease signatures correlated with experimental drug efficacy in breast, liver, and colon cancers [51].

Application of transcriptome-based hit discovery

We briefly present a practical use of transcriptome-based approaches involving the well-established EGFR inhibitor erlotinib for

Please cite this article in press as: Mun, J. et al. A guide for bioinformaticians: 'omics-based drug discovery for precision oncology, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.08.004



(a)

Rank	Cell_treated time_Compound ID_Drug_Target	Trend	P_{adj}
1.	A375_24H_BRD-K16478699_PLX-4720_BRAF	Pos	3.58E^{-55}
2.	A375_24H_BRD-K09951645_dabrafenib_BRAF	Pos	8.85E^{-55}
3.	PC3_24H_BRD-A75409952_wortmannin_PI3K	Pos	1.03E^{-53}
4.	A375_24H_BRD-K57080016_selumetinib_MEK	Pos	1.58E^{-53}
5.	A375_24H_BRD-K16478699_PLX-4720_BRAF	Pos	2.15E^{-53}
6.	A375_24H_BRD-K09951645_dabrafenib_BRAF	Pos	3.11E^{-53}
7.	A375_24H_BRD-U73308409_vemurafenib_BRAF	Pos	5.97E^{-53}
8.	A375_24H_BRD-U73308409_vemurafenib_BRAF	Pos	6.35E^{-53}
9.	HT29_24H_BRD-K41895714_AS-605240_PI3K	Pos	9.43E^{-53}
10.	A375_24H_BRD-K09951645_dabrafenib_BRAF	Pos	4.90E^{-52}
11.	MCF10A_24H_BRD-K41895714_AS-605240_PI3K	Pos	5.05E^{-52}
12.	MCF10A_24H_BRD-K05804044_AZ-628_BRAF	Pos	2.52E^{-51}
13.	MCF10A_24H_CMAP-L030_EGFR ligand	Neg	2.59E^{-51}
76.	PC3_24H_BRD-K64052750_gefitinib_EGFR	Pos	3.94E^{-45}
342.	MCF10A_24H_BRD-K70401845_erlotinib_EGFR	Pos	1.38E^{-38}

(b)

Rank	Cell_treated time_Compound ID_Drug_Target	Trend	P_{adj}
1.	PC3_24H_BRD-K69932463_AZD-8055_mTOR	Neg	1.66E^{-213}
2.	PC3_24H_BRD-K12184916_NVP-BEZ235_PI3K	Neg	2.59E^{-210}
3.	BT20_24H_BRD-K68174511_torin-2_mTOR	Neg	3.49E^{-209}
4.	BT20_24H_BRD-K52911425_GDC-0941_PI3K	Neg	2.56E^{-207}
5.	BT20_24H_BRD-K58772419_AZD-6482_PI3K	Neg	7.83E^{-203}
6.	BT20_24H_BRD-K40175214_torin-1_mTOR	Neg	2.73E^{-202}
7.	BT20_24H_BRD-A45498368_WYE-125132_mTOR	Neg	2.21E^{-189}
8.	H1299_6H_BRD-K03618428_PP-110_PI3K	Neg	7.43E^{-189}
9.	VCAP_48H_BRD-A79768653_sirolimus_mTOR	Neg	4.72E^{-188}
10.	MCF10A_24H_CMAP-L274_TGFa ligand	Pos	7.43E^{-183}
13.	SNGM_6H_BRD-K85606544_neratinib_EGFR/HER2	Neg	3.92E^{-171}
16.	MCF10A_24H_BRD-K70401845_erlotinib_EGFR	Neg	3.79E^{-163}
19.	MCF10A_24H_CMAP-L097_EGF ligand	Pos	2.88E^{-162}
24.	MCF10A_24H_BRD-A58767537_afatinib_EGFR	Neg	2.62E^{-153}
27.	MCF10A_24H_BRD-K08799216_pelitinib_EGFR	Neg	6.14E^{-151}
29.	MDAMB231_24H_BRD-K37687095_AZD-8330_MEK	Neg	2.87E^{-148}
30.	MCF10A_24H_BRD-K64052750_gefitinib_EGFR	Neg	2.64E^{-146}

Drug Discovery Today

FIGURE 2

An example of transcriptome-based hit discovery. **(a)** Drug–drug connectivity between the erlotinib signature and Library of Integrated Network-based Cellular Signatures (LINCS) profiles retrieves compounds with similar transcriptional activity to erlotinib. Pos, positive match; Neg, negative match; P_{adj}, adjusted P-value. **(b)** Disease–drug connectivity between the EGFR-mutant signature and LINCS profiles retrieves candidate inhibitors that reverse disease signatures of EGFR-mutant lung cancer.

EGFR-mutant lung cancer. Drug–drug connectivity can reveal hits with distinct chemical structures but similar biological effects as erlotinib, whereas disease–drug connectivity can identify hits that potentially reverse the disease signature of EGFR-mutant lung cancer (Fig. 2). We used three public data sets: (i) drug signature: HCC827 cells treated with 1 mM erlotinib for 24 h (GSE51212); (ii) drug reference: 473 647 LINCS profiles (GSE92742); and (iii) disease signature: EGFR-mutant lung cancer versus matched normal tissues (GSE40419). The drug–drug connectivity using gcMAP [52] revealed that erlotinib showed significant positive similarity to inhibitors of EGFR, RAF, MEK, and PI3K and negative similarity to EGF in the list of top-ranked hits (Fig. 2a). The disease–drug connectivity revealed inhibitors against EGFR and PI3K-mTOR as the foremost hits for EGFR-mutant lung cancer (Fig. 2b). This analysis identified hits by reconstituting the prior knowledge that EGFR transmits signals from EGF binding through the downstream RAS-RAF-MEK and PI3K-AKT cascades, reflecting the capability of transcriptome-based analyses to identify candidate drugs for the treatment of biological states of interest.

Indeed, transcriptome-based approaches have been used to discover drug candidates in cancer. Applying the approach combined with pathway-based prioritization, Nadine *et al.* repositioned US Food and Drug Administration (FDA)-approved tricyclic antidepressants as inhibitors against lung and neuroendocrine tumors [53]. Vera *et al.* found that citalopram is a therapeutic option that reverses a metastatic signature of colorectal cancer [54]. The Reverse Gene Expression Score using LINCS estimated the drug potency in the reversal of disease signatures, identifying pyrvinium pamoate as a potent hit in liver cancer [51]. As a modified method of transcriptome-based approaches, OncoTreat, with which compounds are prioritized by their ability to reverse cancer dependency driven by master regulators, repurposed the HDAC inhibitor entinostat as a potent drug for neuroendocrine tumors, resulting in the launch of a clinical trial [55]. SynergySeq, another modified approach, retrieves drug combinations by integrating drug signatures and disease signatures [56].

Compound evaluation

Many hits discovered in public data have already been patented. Even if customized libraries are used, discovered hits should be pharmacologically improved to generate lead compounds characterized by higher potency and selectivity, favorable metabolism, and limited toxicity [57]. Given that this lead optimization phase involves iterative rounds of chemical synthesis,

Please cite this article in press as: Mun, J. et al. A guide for bioinformaticians: 'omics-based drug discovery for precision oncology, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.08.004

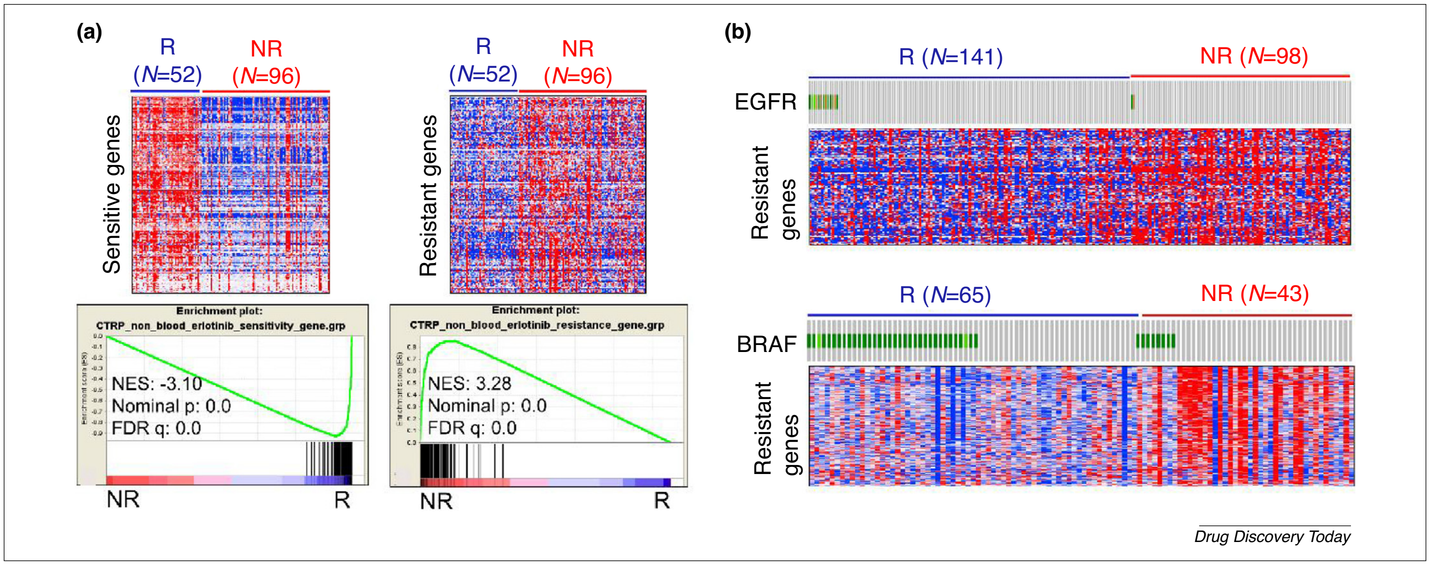


FIGURE 3

An example of transcriptome-based drug sensitivity prediction and biomarker identification. **(a)** Heatmap and Gene Set Enrichment Analysis (GSEA) analyses indicate that the predicted 52 responders (R) and 96 nonresponders (NR) exhibit preferential high-expression trends for CTRP-driven erlotinib-sensitive and erlotinib-resistant genes, respectively. **(b)** Significant enrichment ($P < 0.05$) of *EGFR* and *BRAF* mutations in patient subgroups predicted to be responders against erlotinib and vemurafenib, respectively. Heatmaps exhibit nonresponder-specific high-expression trends for CTRP-driven resistant genes against erlotinib and vemurafenib.

bioinformaticians must collaborate with medicinal chemists to modify chemical motifs and satisfy patentability requirements by inspecting chemical synthesis issues. At this chemistry-first phase, the contribution of 'omics-based approaches might be limited, but producing proper 'omics data upon compound treatment can facilitate evaluations of several aspects of leads. Analytical points for omics-based lead evaluation are described below.

(i) The use of proper data: 'omics data can be selectively produced to reflect the functionality of target proteins. For example, RNA-seq and ChIP-seq revealed that ABBV-744 acts as a BRD4 inhibitor through displacement of the protein from androgen receptor-associated super-enhancers in prostate cancer [58]. Likewise, generating data, such as RNA-seq, phosphoproteome, and KINOMEscan data for kinase inhibitors, and RNA-seq and metabolome data for enzymatic inhibitors, is a practical strategy; (ii) mechanism of action (MoA): the MoA is investigated based on whether drug signatures are enriched compared with any reference gene-set [e.g., Molecular Signatures Database (MSigDB), Gene Ontology, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and Enrichr libraries). Gene-set analyses are performed by Fisher's exact test [Enrichr, DAVID, Ingenuity Pathway Analysis (IPA), GoMiner, MAPPFinder, and EASE] or the Kolmogorov–Smirnov statistic (Gene-Set Enrichment Analysis; GSEA) [59]; (iii) similarity with genetic perturbagens: using genetic perturbation data from LINCS and GEO, the efficacy of compounds can be assessed by comparing their signatures with those of genetic perturbation of the common target and, thus, determining the degree of similarity. Higher similarity indicates a tendency toward more potent and selective leads; and (iv) decision-making for development: the transcriptome is sometimes informative for go/no-go decisions during lead optimization. Previously, the adverse effects of leads were determined by the downregulation of tubulin genes in the development of PDE10A inhibitors and the downregulation of mitochondrial genes in the development of EGFR inhibitors [60].

Prediction of drug responses

During lead optimization, hundreds of compounds are chemically synthesized and committed to a screening campaign by repetitive biochemical and cellular assays. Biochemically potent leads are selected based on IC_{50} and then subjected to cell-based assays that measure the drug sensitivity of cancer cell lines (CCLs) according to the IC_{50}, EC_{50} or area under the curve (AUC) [61]. These repetitive tasks, although mandatory for anticancer drugs, are sometimes labor intensive because hundreds of leads are evaluated across dozens to hundreds of CCLs.

Several computational models, including single-gene predictors, multivariate classifiers, machine learning, and quantitative structure–activity relationships, have been developed to predict drug sensitivity [62]. Transcriptome-based approaches can also be tailored to predict drug sensitivity by assessing the connectivity of three signatures (Fig. 1c): (i) the responder signature, where responders and nonresponders are experimentally identified by IC_{50}, and differential gene expression in responders relative to nonresponders is calculated using the CCLE; (ii) the drug signature; and (iii) the CCL and/or patient signature, where transcriptional signatures of CCLs are generated by comparison with average gene expression values in cell type-matched CCLs, and patient signatures are created by comparison with average gene expression values in normal tissues. Given that CCLs and/or patients positively matched with responder signatures might have responder-like transcriptional programs, they are predicted to be responders (Fig. 1c). CCLs and/or patients inversely matched with drug signatures are also predicted to be responders (Fig. 1c), whereas CCLs and/or patients positively matched with drug signatures are predicted to be nonresponders because drugs might exaggerate cancer transcriptional programs.

Here, we illustrate practical uses of the transcriptome-based method described earlier using two data sets, an erlotinib-induced signature (GSE80344) and the CCLE. Connectivity between these signatures classified 52 lung CCLs as responders (reverse match) and 96 lung CCLs as

Please cite this article in press as: Mun, J. et al. A guide for bioinformaticians: 'omics-based drug discovery for precision oncology, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.08.004

nonresponders (positive match) (Figs. 1c and 3a). This simple transcriptome-based prediction resulted in ~80% accuracy compared with previous experimental results [63]. The result can be also validated using independent CTRP data [64]; CTRP-driven erlotinib-sensitive and erlotinib-resistant genes were exclusively expressed in subgroups predicted to be responders and nonresponders, respectively (Fig. 3a). This example reveals that transcriptome data alone can aid in selecting CCLs for cell-based assays by predicting drug sensitivity.

Biomarker identification

Biomarkers are molecular indicators of drug applicability in individual patients and, therefore, are required for successful clinical trials and precision medicine. The recent recognition that targets, drugs, and biomarkers constitute an indispensable therapeutic triad emphasizes the importance of biomarkers in drug discovery [65].

As evidenced by previous pharmacogenomics studies [Genomics of Drug Sensitivity in Cancer (GDSC) and CTRP] [66,67], 'omics data provide fundamental information for biomarker identification. A study in which 82% of cancer dependencies were predicted by transcriptomes highlights the rationale for transcriptome-based biomarker identification (16% by mutations and 2% by copy number) [14]. The predictive power of mutations might be underestimated because of the low frequency of mutations. Indeed, the administration of many approved drugs is guided by the presence of driver mutations that elicit oncogene addiction [68].

We introduce an 'omics-based workflow to predict biomarkers. For a proof of concept, two drugs with known biomarkers, erlotinib (GSE51212) and vemurafenib (GSE99898), were selected. First, sensitivity to these drugs was predicted by estimating the connectivity between drug signatures and disease signatures from the TCGA (Fig. 1c). Then, comparative multiomics analyses of the predicted responders and nonresponders (Fig. 1d) revealed that EGFR and BRAF mutations, which are known biomarkers for erlotinib and vemurafenib, respectively, were significantly enriched in responders (Fig. 3b). This example shows that a drug signature alone can facilitate the identification of responder-associated biomarkers. However, when transcriptome data are derived from tumor tissues, cancer cell-specific signatures can be confounded by intratumoral heterogeneity in cell type (e.g., immune cells, fibroblasts, and vascular cells) [69]. This issue can be addressed by isolating cancer cell-specific signatures using single-cell RNA-seq or deconvoluting the cell type proportions using cell type-specific transcriptome data [70].

Concluding remarks

Here, we have outlined the bioinformatics contributions during early drug discovery by highlighting the practical applications of 'omics data. Despite the advantages of 'omics-based approaches, their ability is sometimes limited. For example, drug signatures at given time points and/or concentrations might not faithfully represent drug activities. This problem could be complemented by integration with cheminformatics and/or artificial intelligence-based modeling considering the structure and/or pharmacophore properties of targets and compounds [42]. Another limitation is that in vitro 'omics-based activity cannot be extrapolated to in vivo activity, and even in vivo preclinical activity does not guarantee clinical efficacy [48]. Given that an understanding of pharmacokinetics, drug metabolism, and toxicity could be the missing link for this discrepancy [46], the establishment of well-curated data for these in vivo drug properties is mandatory.

This review described the importance of bioinformatics-guided development for both bioinformaticians and other investigators. Investigators accustomed to traditional drug development might remain skeptical regarding bioinformatics-guided development. Nonetheless, with the advent of precision medicine, reorientation toward novel collaborations among multidisciplinary experts for efficient, well-organized drug development is warranted.

Acknowledgments

This study was supported by grants from the Korea Research Institute of Chemical Technology (KK2031-30) and the Korea Institute of Science and Technology Information (K-20-L02-C10-S01).

References

1 Bailey, M.H. *et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell* 173 (2), 371–385
2 Gaudelli, N.M. *et al.* (2017) Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* 551 (7681), 464–471
3 Nelson, M.R. *et al.* (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet* 47 (8), 856–860
4 Hingorani, A.D. *et al.* (2019) Improving the odds of drug development success through human genomics: modelling study. *Sci. Rep.* 9 (1), 18911
5 Dopazo, J. (2014) Genomics and transcriptomics in drug discovery. *Drug Discov. Today* 19 (2), 126–132
6 Pon, J.R. and Marra, M.A. (2015) Driver and passenger mutations in cancer. *Annu. Rev. Pathol.* 10, 25–50
7 Lawrence, M.S. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499 (7457), 214–218
8 Vogelstein, B. *et al.* (2013) Cancer genome landscapes. *Science* 339 (6127), 1546–1558
9 Gerlinger, M. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366 (10), 883–892
10 Jolly, C. and Van Loo, P. (2018) Timing somatic events in the evolution of cancer. *Genome Biol.* 19 (1), 95
11 Malikic, S. *et al.* (2019) Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat. Commun.* 10 (1), 2750
12 McLaren, W. *et al.* (2016) The Ensembl variant effect predictor. *Genome Biol.* 17 (1), 122
13 Dang, C.V. *et al.* (2017) Drugging the 'undruggable' cancer targets. *Nat. Rev. Cancer* 17 (8), 502–508
14 Tsherniak, A. *et al.* (2017) Defining a cancer dependency Map. *Cell* 170 (3), 564–576 e516
15 Luo, J. *et al.* (2009) Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell* 136 (5), 823–837
16 Kuleshov, M.V. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44 (W1), W90–W97
17 Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102 (43), 15545–15550
18 Cai, M. and Li, L. (2017) Subtype identification from heterogeneous TCGA datasets on a genomic scale by multi–view clustering with enhanced consensus. *BMC Med. Genomics* 10 (Suppl. 4), 75
19 van Dam, S. *et al.* (2018) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform.* 19 (4), 575–592
20 Vasaikar, S. *et al.* (2019) Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* 177 (4), 1035–1049
21 Jiang, Y. *et al.* (2019) Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* 567 (7747), 257–261
22 Butler, M. *et al.* (2020) MGMT status as a clinical biomarker in glioblastoma. *Trends Cancer* 6 (5), 380–391
23 Ong, C.T. and Corces, V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet* 12 (4), 283–293
24 Sengupta, S. and George, R.E. (2017) Super-enhancer-driven transcriptional dependencies in cancer. *Trends Cancer* 3 (4), 269–281
25 Roadmap Epigenomics, C. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518 (7539), 317–330
26 Lim, B. *et al.* (2016) Genomic and epigenomic heterogeneity in molecular subtypes of gastric cancer. *World J. Gastroenterol.* 22 (3), 1190–1201
27 Cancer Genome Atlas Research, N (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513 (7517), 202–209
28 Patti, G.J. *et al.* (2012) Innovation: metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* 13 (4), 263–269
29 Wishart, D.S. (2016) Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev Drug Discov.* 15 (7), 473–484

Please cite this article in press as: Mun, J. et al. A guide for bioinformaticians: 'omics-based drug discovery for precision oncology, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.08.004

30 Luck, K. *et al.* (2017) Proteome-scale human interactomics. *Trends Biochem. Sci.* 42 (5), 342–354
31 Kanhaiya, K. *et al.* (2017) Controlling directed protein interaction networks in cancer. *Sci. Rep.* 7 (1), 10327
32 Liu, L. *et al.* (2016) Modeling co-occupancy of transcription factors using chromatin features. *Nucleic Acids Res.* 44 (5), e49
33 Guo, J. *et al.* (2019) Degrading proteins in animals: 'PROTAC'tion goes *in vivo*. *Cell Res.* 29 (3), 179–180
34 Koscielny, G. *et al.* (2017) Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* 45 (D1), D985–D994
35 Carvalho-Silva, D. *et al.* (2019) Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* 47 (D1), D1056–D1065
36 Rouillard, A.D. *et al.* (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)* 2016, baw100
37 Nguyen, D.T. *et al.* (2017) Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* 45 (D1), D995–D1002
38 McDonald, E.R., 3rd *et al.* (2017) Project DRIVE: a compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. *Cell* 170 (3), 577–592
39 McFarland, J.M. *et al.* (2018) Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* 9 (1), 4610
40 Kessler, J.D. *et al.* (2012) A SUMOylation-dependent transcriptional subprogram is required for Myc-driven tumorigenesis. *Science* 335 (6066), 348–353
41 Ghandi, M. *et al.* (2019) Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 569 (7757), 503–508
42 Schneider, P. *et al.* (2020) Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* 19 (5), 353–364
43 Cui, W. *et al.* (2020) Discovering anti-cancer drugs via computational methods. *Front Pharmacol.* 11, 733
44 Zhu, T. *et al.* (2013) Hit identification and optimization in virtual screening: practical recommendations based on a critical literature analysis. *J. Med. Chem.* 56 (17), 6560–6572
45 Lyu, J. *et al.* (2019) Ultra-large library docking for discovering new chemotypes. *Nature* 566 (7743), 224–229
46 Zheng, M. *et al.* (2013) Computational methods for drug design and discovery: focus on China. *Trends Pharmacol. Sci.* 34 (10), 549–559
47 Lamb, J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313 (5795), 1929–1935
48 Huang, C.T. *et al.* (2018) A large-scale gene expression intensity-based similarity metric for drug repositioning. *iScience* 7, 40–52
49 Subramanian, A. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171 (6), 1437–1452
50 Qu, X.A. and Rajpal, D.K. (2012) Applications of Connectivity Map in drug discovery and development. *Drug Discov. Today* 17 (23–24), 1289–1298
51 Chen, B. *et al.* (2017) Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat. Commun.* 8, 16022
52 Sandmann, T. *et al.* (2014) gCMAP: user-friendly connectivity mapping with R. *Bioinformatics* 30 (1), 127–128
53 Jahchan, N.S. *et al.* (2013) A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discov.* 3 (12), 1364–1377
54 van Noort, V. *et al.* (2014) Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. *Cancer Res.* 74 (20), 5690–5699
55 Alvarez, M.J. *et al.* (2018) A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nat. Genet* 50 (7), 979–989
56 Stathias, V. *et al.* (2018) Drug and disease signature integration identifies synergistic combinations in glioblastoma. *Nat. Commun.* 9 (1), 5315
57 Hughes, J.P. *et al.* (2011) Principles of early drug discovery. *Br. J. Pharmacol.* 162 (6), 1239–1249
58 Faivre, E.J. *et al.* (2020) Selective inhibition of the BD2 bromodomain of BET proteins in prostate cancer. *Nature* 578 (7794), 306–310
59 Tryputsen, V. *et al.* (2014) Using Fisher's method to identify enriched gene sets. *Stat. Biopharm. Res.* 6 (2), 154–162
60 Verbist, B. *et al.* (2015) Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project. *Drug Discov. Today* 20 (5), 505–513
61 Pozdeyev, N. *et al.* (2016) Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. *Oncotarget* 7 (32), 51619–51625
62 Nguyen, L. *et al.* (2016) Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. *F1000Res* 5 ISCB Comm J 2927
63 Coldren, C.D. *et al.* (2006) Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. *Mol. Cancer Res.* 4 (8), 521–528
64 Rees, M.G. *et al.* (2016) Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* 12 (2), 109–116
65 McMillan, E.A. *et al.* (2018) Chemistry-first approach for nomination of personalized treatment in lung cancer. *Cell* 173 (4), 864–878
66 Garnett, M.J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483 (7391), 570–575
67 Iorio, F. *et al.* (2016) A landscape of pharmacogenomic interactions in cancer. *Cell* 166 (3), 740–754
68 Twomey, J.D. *et al.* (2017) Drug-biomarker co-development in oncology - 20 years and counting. *Drug Resist Updat* 30, 48–62
69 Schmidt, F. and Efferth, T. (2016) Tumor heterogeneity, single–cell sequencing, and drug resistance. *Pharmaceuticals (Basel)* 9, 33
70 Newman, A.M. *et al.* (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12 (5), 453–457

Please cite this article in press as: Mun, J. et al. A guide for bioinformaticians: 'omics-based drug discovery for precision oncology, Drug Discov Today (2020), https://doi.org/10.1016/j.drudis.2020.08.004