6.2 Bridging Information on Drugs, Targets, and Diseases

Andreas Steffen and Bertram Weiss

Bayer Pharma Aktiengesellschaft, PH-DD-TRG-CIPL-Bioinformatics, Müllerstr. 178, Berlin 13342, Germany

Learning Objectives

- To discuss the importance of data curation and integration for enabling effective data analysis.
- To describe how integrative data analysis using proprietary and public data sources impacts pharma R&D along the value chain.

Outline

6.2.1 Introduction

The pharmaceutical industry's core mission is to develop novel, innovative, efficacious, and at the same time cost-effective therapies [1]. In order to move beyond best-in-class drugs, pharmaceutical companies strive for first-in-class drugs and are in high need for new strategies that include external innovation in academia, precompetitive collaborations, biotech investments, in-licensing or acquisitions, and fostering innovation within the company's own R&D. One key strategy and competitive factor is the capability to maximally exploit the ever-growing life science data.

While traditionally new drug targets often emerged from textbook knowledge, biological reasoning, me-too approaches, or even serendipity, pharmaceutical companies now emphasize the importance of investing into innovative ways to identify novel and promising therapeutic concepts. The discipline of computational life sciences has matured and now impacts pharma R&D along the entire value chain. For bioinformatics target discovery, biomarker discovery, and indication expansion are three major fields, whereas chemoinformatics is focusing on library design, lead finding, and optimization. More data-driven and systematic approaches to drug and target discovery shall move pharma R&D beyond the often cited serendipity of the past [2].

Analysis of existing large experimental datasets coming from both internal and mainly external sources enable us today to quickly formulate actionable hypotheses that guide experimentalists to focus on the most promising

Applied Chemoinformatics: Achievements and Future Opportunities, First Edition.
Edited by Thomas Engel and Johann Gasteiger.

experiments. However, the persistent rigid data silos and the often unstructured non-integrated nature of the data hamper data scientists to effectively leverage the large amount of existing data. Ideally this requires analysis-ready data so that computational scientists can derive hypothesis with minimal technical barriers. In this way most time is spent on identifying and algorithmically formulating the biological questions and their in-depth analysis and not on data cleansing and curation.

In the following sections we introduce key data sources for compounds, targets, and diseases (see Table 6.2.1), which are the three most relevant data entities for preclinical data science. Pharma R&D data is loosely structured around these three entities. Repositories do exist for all three; however, of particular strong interest are evidence data that relate these entities. Within the last decade new data repositories emerged that integrate the relations and provide evidences for them, for example, opentargets.org [3], ChEMBL [4], drug2gene [5], or PhenomicDB [6].

We refer our readers to the original publications for details and rather exemplify and discuss here the integrative uses of these data sources for the main bio- and chemoinformatics activities in target identification, patient stratification, indication expansion, and toxicity prediction. The discussion provides a perspective where we see future developments in the field.

6.2.2 Existing Data Sources

For key data sources of compounds, targets, and diseases please see Table 6.2.1.

6.2.3 Drug Discovery Use Cases in Computational Life Sciences

6.2.3.1 Target Mining

A healthy pipeline of a pharmaceutical company requires a constant flow of promising new targets. Bioinformaticians are asked to deliver new drug target proposals relevant either for so far unmet medical indications or for improving the therapeutic options of already but not optimally treated diseases. Coming up with new concepts to identify innovative new targets is of high importance for pharmaceutical companies as this can provide a competitive advantage in the search for new first-in-class drugs. In oncology, we are facing a comparably comfortable situation in terms of availability of genomic patient data that can be mined and exploited to identify novel druggable targets, whereas in most other indications this data avalanche is somewhat less pronounced [7].

New target identification campaigns might, for example, start with a question: what are frequently mutated genes in lung cancer that act as oncogenes and that could be targeted via small molecules? Mutation data of a large number of cancer patients can, for example, be obtained from the TCGA database [8]. The functional impact of a mutation has to be assessed in detail as gain-of-function

Table 6.2.1 Selected key data resources focusing on the three pharma R&D relevant entities: targets, compounds, and diseases.

Name	Short description	URL (January 2018)
Target-centric repositories		
NCBI Gene	All information about genes	ncbi.nlm.nih.gov/gene
Ensembl	All information about genes	www.ensembl.org
UniProt	All information about proteins	www.uniprot.org
Reactome	Pathway database	www.reactome.org
WikiPathways	Pathway database with public curation	www.wikipathways.org
Pathway Commons	Pathway meta-database	www.pathwaycommons.org
KEGG	Pathway database	www.genome.jp/kegg
Compound-centric repositories		
ChemSpider	Compound properties	www.chemspider.com
ZINC	Compound structures prepared for cheminformatics studies	zinc.docking.org
eMolecules	Compound ordering	www.emolecules.com
Disease-/phenotype-centric repositories		
OMIM	Catalogue of Mendelian diseases	www.ncbi.nlm.nih.gov/omim
Compound–target-centric repositories		
ChEMBL	Repository for biochemical activities	www.ebi.ac.uk/chembl
PubChem	Repository for biochemical activities	pubchem.ncbi.nlm.nih.gov
drug2gene	Meta-database for compound–target relations	
ChemBank	Repository for biochemical activities	chembank.broadinstitute.org
Compound–disease/phenotype-centric repositories		
DrugBank	Repository for drugs and their indications	
ClinicalTrials	Information about clinical trials	clinicaltrials.gov
Achilles	RNAi screening in cancer cell lines	portals.broadinstitute.org/achilles
CTD2	Compound informer set on cancer cell lines	ctd2.nci.nih.gov
Target–disease-centric repositories		
PhenomicDB	Meta-database for phenotype-disease relations	
opentargets	Repository around evidences for target–disease associations	www.opentargets.org
GWAS Catalog	Catalogue SNP to disease relations	www.ebi.ac.uk/gwas
HuGe Navigator	Gene-disease associations	www.cdc.gov/genomics/hugenet/hugenavigator.htm
CCLE	Cancer cell line encyclopedia	portals.broadinstitute.org/ccle
TCGA	Genomic data of cancer patients	tcga-data.nci.nih.gov

mutations can be directly targeted, whereas loss-of-function mutations that are more likely to occur in tumor suppressors are harder to address therapeutically. Many somatic genetic alterations are in fact due to increased genomic instability in tumors and contrary to so-called driver events do not contribute to tumor development [9]. These events are called passenger alterations and should usually be filtered out as they are not causal to tumor development. A range of bioinformatics algorithms such as MutSigCV, mutation assessor, or IntOGen are thus employed to assess the significance and the functional impact of a mutation [10–12]. One interesting and integrated approach is to look for spatial clusters of somatic mutations in order to identify functional hot spots that could point to a gain-of-function alteration [13–15].

Within the Achilles project, genome-wide RNAi experiments are conducted in a large number of cancer cell lines in order to uncover targetable vulnerabilities of tumors [16]. In contrast, the CTD2 project makes use of a so-called informer set of small molecules with known mode of action and provides experimentally determined sensitivity of a large number of cancer cell lines [17]. Such perturbator datasets can provide first insights into the phenotype of a gene's inhibition and thus are useful resources in the quest for new cancer targets.

Nowadays, text mining is a powerful tool to provide literature evidence for a gene's relevance to a disease of interest helping to quickly provide context of already existing knowledge about involvement of a gene in a disease and is especially useful for larger lists of candidate genes [18, 19].

Further, annotations such as the expression of the gene in the targeted tumor type and in the human body in general [20, 21], the availability of crystal structures in the PDB [22], disease–target association scores based on genetic linkage or genetic studies in model organisms [3, 23], the druggability of a target [24, 25], or competitor information from business intelligence databases can help to narrow down candidate lists into manageable sizes.

Researchers prioritize long lists of targets in a multiparametric fashion considering all these data and literature in order to prioritize drug target candidates systematically in a data-driven fashion and ultimately propose targets that will hopefully lead to new therapeutic options to the benefit of the patients. Integrative platforms to support this process exist and provide very effective ways to obtain target insights that support the target prioritization process [3, 26].

While the previous section provided a description of a target-centric approach to mine for new molecular entities to treating diseases, a comparably large part of new chemical starting points stems from so-called phenotypic screens, in which a desired phenotypic change of a cellular systems is screened for [27]. This is especially helpful to address pathways without suitable drug targets [28]. However, in many cases the phenotype-causing targets of the resulting screening hits are unknown and thus have to be elucidated with significant experimental efforts [29–31]. *In silico* target fishing approaches can be applied to focus experimental efforts by narrowing down the number of possible targets that could cause the phenotypic readout [30, 32, 33]. These approaches make use of large bioactivity data sources such as ChEMBL [34], ChemBank [35], or drug2gene [5] and the vast amounts of bioactivity data within pharmaceutical companies.

6.2.3.2 Biomarker Search for Patient Stratification

Once drug candidates are identified, preclinical research tries to develop hypotheses that patient populations could maximally benefit from a drug [36]. A biomarker identifying those patients has become a key value driver for successful drug development nowadays [37, 38]. In preclinical R&D often larger cell line sensitivity panels are performed. The Cancer Cell Line Encyclopedia (CCLE) provides genomic annotations of more than 1000 cell lines [40]. Further genomically characterized cell lines can be found in the Sanger cell line project and in the Genentech dataset [40, 41]. Differential sensitivities of drugs on these cell lines can then be related to genomic alterations of the cell lines via a range of algorithmic approaches, and predictive biomarker candidates for sensitivity can possibly be derived. The search for such biomarkers often starts with univariate approaches, to see whether simple and clinically exploitable markers for sensitivity exist, for example, the mutation or the copy number status of a known frequently altered gene. Multivariate approaches such as elastic net try to identify combinations of alterations that predict the sensitivity toward the compound [39, 42]. Often differential gene expression between the sensitive and the resistant cell lines are calculated [43–45]. Differentially expressed genes can then be further analyzed in terms of underlying biology, for example, the enrichment of genes from a signaling pathway or a known subtype of cancer. Tools such as gene set enrichment analysis [46] or the online platform DAVID [47] can be of great help to better understand the biology that causes certain patient populations to benefit from a drug and vice versa. Again, perturbator datasets such as Achilles or CTD2 can be useful resources for biomarker identification and might help to find genetic alterations that are predictive for sensitivity toward a drug-mediated inhibition of a target [16, 17].

While cellular disease models have proven useful to uncover genetic markers as predictors of drug sensitivity in cell lines, it is more and more understood that they can substantially deviate from tumors [48], and thus hypotheses derived from such models might not be translatable to the clinics [49]. New approaches try to improve the translatability of preclinical findings by establishing genomically characterized patient-derived primary cancer cells [50, 51] or patient-derived xenografts [52] and studying drug sensitivity on these models that are more likely to reflect the biology of real tumors.

6.2.3.3 Toxicity Prediction

Chemogenomic resources combined with sophisticated algorithmic approaches can be very useful to uncover reasons for compound-mediated toxicities or side effects. As described previously, based on the vast amounts of bioactivity data in publically available data sources and in-house repositories of pharmaceutical companies alike, target activity prediction models can be calculated. A range of toxicities can often be attributed to so-called off-target effects, and thus the set of the corresponding activity models can be applied to computationally predict target-mediated toxicities of compounds. In a recent study, Lounkine *et al*. presented a computational method termed similarity ensemble approach (SEA) to predict the activity of drugs on 73 unintended "side effect" targets

[53]. This method predicts whether a drug molecule could bind to an unwanted target based on the chemical similarity of the drug and the ensemble of all known binders of the unwanted target and a statistical model that controls for random similarity [54]. About half of the computational predictions could be experimentally verified, underlining the power of the combination of sophisticated algorithms and curated data to support drug discovery projects. Promising new algorithmic approaches that can digest large data resources might help to further positively impact drug discovery projects. In a recent FDA-initiated prospective competition termed Tox21, algorithmic groups were asked to predict toxic effects for 647 compounds using a training set comprising toxicity-compound associations for more than 11,000 chemical compounds [55]. Among all submissions, Mayr *et al*. ranked top by using a multitask deep learning approach [56].

While we believe that algorithmic innovations are key to intelligently mine existing data resources, well-curated data resources will always be much sought after and often remain the bottleneck. The consortium-based eTOX project aims to develop a semantically embedded database comprising industry legacy data and public toxicology data [57]. This precompetitive initiative allows for the mining of data from competitors using a broker model, so that ultimately all participating units benefit to bring safer compounds to the patient.

6.2.3.4 Indication Expansion

Once a candidate drug has shown to be efficacious and safe in animals and has thus been selected for clinical advancement, pharmaceutical companies try to identify further new upside indications besides the originally intended core indications [58]. In a first step, text and database mining and careful manual inspection is applied to reveal further target–disease associations mainly based on the concept of diseases sharing similar pathological mechanisms (e.g., inflammation in rheumatoid arthritis, endometriosis, or psoriasis; fibrosis in lung fibrosis, liver cirrhosis, or tissue scarring; angiogenesis in cancer, age-related macular disease, or endometriosis) that eventually might point to new patient populations. A new, well-integrated resource for target–disease relationships is opentargets.org [3].

Another promising approach to come to new indications is based on the Connectivity Map concept. In their landmark publication, Lamb *et al*. describe a systematic method to uncover functional connections of diseases, genes, and drugs [59]. This approach has been applied by pharmaceutical companies for indication expansion [60, 61]. Gene expression profile changes induced by a compound are compared with gene expression changes in diseased versus normal tissues (so-called gene signatures) in order to eventually identify those diseases whose gene expression change could possibly be counter-regulated by the compound. The hypothesis is that if a drug downregulates genes that are upregulated in a disease state (and vice versa), it is possible that patients with this disease could benefit from this drug as it *reverses* the transcriptomic changes of the disease mechanism. The antiepileptic drug topiramate was repositioned for inflammatory bowel disease (IBD). It was a significant hit when gene signatures of 164 drugs were compared with public gene signatures of IBD for anticorrelation [62].

In modern drug discovery it has become a value driver to exploit the potential of a compound in several diseases, and therefore it is a common task for computational groups to evaluate the upside potential of a target already during target identification.

6.2.4 Discussion and Outlook

Preclinical research is positively impacted by effective computational usage of relevant data sources. However, still much valuable internal data in pharmaceutical companies resides in rigid data silos lacking harmonization/standardization, common identifier spaces, or rigorously applied ontologies. It is often difficult to identify relevant data, to access it, and to overcome complex data structures. This severely hampers intelligent and efficacious data analysis. In recent years, many companies spent substantial efforts to enable researchers to creatively mine these data. They have understood that data need to be easily accessible, well integrated, and very importantly analysis ready so that most energy goes into the analyses and the interpretation as such. At the same time, publically available data sources described in Table 6.2.1 increase rapidly and provide substantial means for data-driven drug discovery in industry and academia alike. Easy and sustainable integration of public and internal data is a constant and costly challenge.

While the existence of data is key, it is the intelligent curation and integration that will make algorithmic mining possible. For example, if cellular models for diseases shall be selected, disease descriptions of both patient data and the corresponding cell line models need to be standardized so that they can be mapped onto each other. It is important that the corresponding entities in different data sources are described by standardized identifiers, for example, targets via NCBI Gene or Ensembl identifiers, so that integrative studies are not hampered by the necessity to investigate identifier mappings over and over again. For more complex entities like diseases, use of ontologies is a must. Currently, they are not consistently applied, and for diseases there are many ontologies and none is prevailing. Public funding organizations and precompetitive initiatives, for example, the Pistoia Ontologies Mapping project, start addressing some of these bottlenecks [63]. It is surprising that funding for generating data is readily available but often without considering dedicated budgets for proper dissemination to exploit it.

Whereas our chapter focused mainly on the organization and exploitation of the parts list of a cell and how molecules interact with each other, we have spared out the field of systems biology, where the approach is to understand all the parts and interactions as a whole. Although that is the ultimate goal of course, to our knowledge this Holy Grail has not yet shown to fully deliver sufficiently reliable results to significantly impact drug discovery projects. The foreseeable advent of wearable devices [64] and the digitalization of electronic health records [65] will produce very large amounts of phenotypic/clinical data and provide enormous computational opportunities and challenges. We foresee a future where we have common identifier spaces and ontologies for important entities (genes,

drugs, diseases) and new datasets are semantically annotated right away easing integration. Analysis algorithms are registered and automatically applied to meaningful data. If statistically significant results are obtained, researchers who have registered for topics are alerted. By this, interesting new hypotheses are automatically generated, forwarded, and thus put under consideration to our skilled scientists.

Essentials

- Proper curation and integration of relevant data sources is key to effective data analysis in life sciences.
- Integrative data analysis requires common identifier spaces across datasets so that they can easily be joined and mined.
- Developing data standards, controlled vocabularies, and ontologies is important to cope with the rapidly growing scientific data across industry and academia.

Available Software and Web Services (accessed January 2018)

- http://www.opentargets.org
- http://www.cbioportal.org/
- http://www.ebi.ac.uk/chembl
- http://www.etoxproject.eu/

Selected Reading

- LaMattina, J.L. (2008) *Drug Truths: Dispelling the Myths about Pharma R&D*, John Wiley & Sons, Inc., Hoboken, NJ, 156 pp.
- Lengauer, T. (2008) *Bioinformatics – From Genomes to Therapies*, Wiley-VCH Verlag GmbH, Weinheim, Germany, 1814 pp.
- McKinney, W. (2012) *Python for Data Analysis*, O'Reilly Media, Sebastopol, CA, 463 pp.
- Wilkinson, M.D. *et al* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**. doi: 10.1038/sdata.2016.18

References

[1] Wang, L., Plump, A., and Ringel, M. (2015) *Drug Discovery Today*, **20**, 361–370.

[2] Kubinyi, H. (2006) Ernst Schering Research Foundation Workshop, www.kubinyi.de/schering58-2006.pdf (accessed January 2018).

[3] Opentargets https://www.targetvalidation.org/ (accessed January 2018).

[4] Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., and Davies, M. Krüger, F.A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., Overington, J.P. (2014) *Nucleic Acids Res.*, **42**, D1083–D1090.
[5] Roider, H.G., Pavlova, N., Kirov, I., Slavov, S., Slavov, T., Uzunov, Z., and Weiss, B. (2014) *BMC Bioinf.*, **15**, 68.
[6] Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H.D., and Weiss, B. (2007) *Nucleic Acids Res.*, **35**, D696–D699.
[7] Gnad, F., Doll, S., Manning, G., Arnott, D., and Zhang, Z. (2015) *BMC Genomics*, **16** (Suppl. 8), S5.
[8] NIH Genomic Data Commons Data Portal, https://gdc-portal.nci.nih.gov/ (accessed January 2018).
[9] Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. Jr.,, and Kinzler, K.W. (2013) *Science*, **339**, 1546–1558.
[10] Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., and Sivachenko, A. (2013) *Nature*, **499**, 214–218.
[11] Sander, C., Schultz, N., Reva, B., Antipin, Y., and Sander, C. (2011) *Nature Commun.*, **39**, e118.
[12] Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., and Jene-Sanz, A. (2013) *Nat. Methods*, **10**, 1081–1082.
[13] Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., and Golub, T.R. (2015) *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E5486–E5495.
[14] Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015) *PLoS Comput. Biol.*, **11**, e1004518.
[15] Engin, H.B., Hofree, M., and Carter, H. (2015) *Pac. Symp. Biocomput.*, 84–95.
[16] Cowley, G.S., Weir, B.A., Vazquez, F., Tamayo, P., Scott, J.A., and Rusin, S. (2014) *Sci. Data*, **1**, 140035.
[17] The Cancer Target Discovery and Development Network (2016) *Mol. Cancer Res.*, **14**, 675–682.
[18] Liu, Y., Liang, Y., and Wishart, D. (2015) *Nucleic Acids Res.*, **43**, W535–W542.
[19] Huang, C.C. and Lu, Z. (2016) *Briefings Bioinf.*, **17**, 132–144.
[20] Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., and Buia, S.A. (2015) *Biopreserv. Biobanking*, **13**, 311–319.
[21] Stokoe, D., Modrusan, Z., Neve, R.M., de Sauvage, F.J., Settleman, J., and Seshagiri, S. (2015) *Nat. Biotechnol.*, **43**, D1113–D1116.
[22] Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., and Burkhardt, K. (2002) *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, **58**, 899–907.
[23] Groth, P., Leser, U., and Weiss, B. (2011) *Methods Mol. Biol.*, **760**, 159–173.
[24] Hopkins, A.L. and Groom, C.R. (2002) *Nat. Rev. Drug Discovery*, **1**, 727–730.
[25] Griffith, M., Griffith, O.L., Coffman, A.C., Weible, J.V., McMichael, J.F., and Spies, N.C. (2013) *Nat. Methods*, **10**, 1209–1210.
[26] Campbell, S.J., Gaulton, A., Marshall, J., Bichko, D., Martin, S., Brouwer, C., and Harland, L. (2010) *Drug Discovery Today*, **15**, 3–15.

[27] Swinney, D.C. and Anthony, J. (2011) *Nat. Rev. Drug Discovery*, **10** (7), 507–519.
[28] McMillan, M. and Kahn, M. (2005) *Drug Discovery Today*, **10**, 1467–1474.
[29] Lee, J. and Bogyo, M. (2013) *Curr. Opin. Chem. Biol.*, **17**, 118–126.
[30] Schirle, M. and Jenkins, J.L. (2016) *Drug Discovery Today*, **21**, 82–89.
[31] Wagner, B.K. and Schreiber, S.L. (2016) *Cell Chem. Biol.*, **23**, 3–9.
[32] Nettles, J.H., Jenkins, J.L., Bender, A., Deng, Z., Davies, J.W., and Glick, M. (2006) *J. Med. Chem.*, **49**, 6802–6810.
[33] Nidhi, Glick, M., Davies, J.W., and Jenkins, J.L. (2006) *J. Chem. Inf. Model.*, **46**, 1124–1133.
[34] Papadatos, G., Gaulton, A., Hersey, A., and Overington, J.P. (2015) *J. Comput.-Aided Mol. Des.*, **29**, 885–896.
[35] Seiler, K.P., George, G.A., Happ, M.P., Bodycombe, N.E., Carrinski, H.A., and Norton, S. (2008) *Nucleic Acids Res.*, **36**, D351–D359.
[36] Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., and Schubert, M. (2016) *Cell*, **166**, 740–754.
[37] Plenge, R.M. (2016) *Sci. Transl. Med.*, **8**, 349ps15.
[38] Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., and Shen, Y. (2015) *Nat. Genet.*, **47**, 856–860.
[39] Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., and Kim, S. (2012) *Nature*, **483**, 603–607.
[40] COSMIC Sanger Cell Line Project, http://cancer.sanger.ac.uk/cell_lines (accessed January 2018).
[41] Klijn, C., Durinck, S., Stawiski, E.W., Haverty, P.M., Jiang, Z., and Liu, H. (2015) *Nat. Biotechnol.*, **33**, 306–312.
[42] Chen, B.J., Litvin, O., Ungar, L., and Pe'er, D. (2015) *PLoS One*, **10**, e0133850.
[43] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015) *Nucleic Acids Res.*, **43**, e47.
[44] Love, M.I., Huber, W., and Anders, S. (2014) *Genome Biol.*, **15** (12), 550.
[45] Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) *Bioinformatics*, **26**, 139–140.
[46] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., and Gillette, M.A. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
[47] da Huang, W., Sherman, B.T., and Lempicki, R.A. (2009) *Nat. Protoc.*, **4**, 44–57.
[48] Domcke, S., Sinha, R., and Levine, D.A. (2013) *Nat. Commun.*, **4**, 2126.
[49] Lieu, C.H., Tan, A.C., Leong, S., Diamond, J.R., and Eckhardt, S.G. (2013) *J. Nat. Cancer Inst.*, **105**, 1441–1456.
[50] Pemovska, T., Kontro, M., Yadav, B., Edgren, H., Eldfors, S., and Szwajda, A. (2013) *Cancer Discovery*, **3** (12), 1416–1429.
[51] Pemovska, T., Johnson, E., Kontro, M., Repasky, G.A., Chen, J., and Wells, P. (2015) *Nature*, **519**, 102–105.
[52] Gao, H., Korn, J.M., Ferretti, S., Monahan, J.E., Wang, Y., and Singh, M. (2015) *Nat. Med.*, **21**, 1318–1325.
[53] Lounkine, E., Keiser, M.J., Whitebread, S., Mikhailov, D., Hamon, J., and Jenkins, J.L. (2012) *Nature*, **486**, 361–367.

[54] Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., and Hufeisen, S.J. (2009) *Nature*, **462**, 175–181.
[55] Tox21 Tox21 Data Browser, https://tripod.nih.gov/tox21/ (accessed January 2018).
[56] Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016) *Front. Environ. Sci.*, **3** (80).
[57] Sanz, F., Carrio, P., Lopez, O., Capoferri, L., Kooi, D.P., and Vermeulen, N.P. (2015) *Mol. Inf.*, **34**, 477–484.
[58] Nielsch, U., Schafer, S., Wild, H., and Busch, A. (2007) *Drug Discovery Today*, **12**, 1025–1031.
[59] Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., and Wrobel, M.J. (2006) *Science*, **313**, 1929–1935.
[60] Cheng, J., Yang, L., Kumar, V., and Agarwal, P. (2014) *Genome Med.*, **6**, 540.
[61] Sirota, M., Dudley, J.T., Kim, J., Chiang, A.P., Morgan, A.A., and Sweet-Cordero, A. (2011) *Sci. Transl. Med.*, **3**, 96ra77.
[62] Dudley, J.T., Sirota, M., Shenoy, M., Pai, R.K., Roedder, S., and Chiang, A.P. (2011) *Sci. Transl. Med.*, **3** (96), 96ra76.
[63] Barnes, M.R., Harland, L., Foord, S.M., Hall, M.D., Dix, I., and Thomas, S. (2009) *Nat. Rev. Drug Discovery*, **8**, 701–708.
[64] Gay, V. and Leijdekkers, P. (2015) *J. Med. Internet Res.*, **17**, e260.
[65] Jensen, A.B., Moseley, P.L., Oprea, T.I., Ellesoe, S.G., Eriksson, R., and Schmock, H. (2014) *Nat. Commun.*, **5**, 4022.