Interesting Properties of Profile Data Analysis in the Understanding and Utilization of the Effects of Drugs

Tadahaya Mizuno,* Katsuhisa Morita, and Hiroyuki Kusuhara*

Graduate School of Pharmaceutical Sciences, the University of Tokyo; Bunkyo-ku, Tokyo 113–0033, Japan.

Received March 31, 2020

Profile data is defined as data which describes the properties of an object. Omics data of a specimen is profile data because its comprehensiveness supports the idea that omics data is numeric information which reflects biological information of the specimen. In general, omics data analysis utilizes an existing body of biological knowledge, while some profile data analysis methods are independent of existing knowledge, which is suitable for uncovering unidentified aspects of a specimen of interest. The effects of a small compound, such as drugs, are multiple, and include unrecognized effects, even by the developers. To uncover such unrecognized effects, it is useful to employ profile data analysis independent of existing knowledge. In this review, we summarize what profile data is, properties of profile data analysis, and current applications of profile data in order to understand and utilize the effects of small compounds, in particular, in a recently developed method to decompose multiple effects of a drug.

Key words profile data analysis; decomposition analysis; factor analysis; orthogonal linear separation analysis (OLSA); connectivity map; principal component analysis

1. INTRODUCTION

One of the main advantages of macromolecular drugs is their high specificity for targets, which reduces the frequency of adverse events even their developers do not expect. In this regard, small molecular drugs are inferior to macromolecular ones, since small molecular drugs can interact with multiple cellular proteins, and such interactions have the potential to induce multiple effects and to cause adverse events. The burdens of adverse events have negative effects not only on patients, but on a wide range of related fields, such as the pharmaceutical industry—one driver of changing trends in drug discovery.1)

However, such unrecognized features of small molecular drugs may turn into an advantage: drug repurposing. For instance, memantine, an anti-influenza agent, has been revealed to act as an antagonist of the N-methyl-D-aspartate receptor, and as such could become a medication to treat dementia.2) Sodium phenylbutyrate, a drug for treating urea cycle disorder, was recently reported to improve the symptoms of benign recurrent intrahepatic cholestasis type 2.3) Such expansion of medication roles utilizes the unrecognized aspects of small molecular drugs, a feature nearly impossible with macromolecular drugs.

In the last two decades, the research environment surrounding the life sciences has greatly changed. A data-driven approach was born, providing a counterpart to the more conventional hypothesis-driven approach based on the existing body of knowledge, and is rapidly growing in parallel with technological innovations in data acquisition methods and computer power. Moreover, recent developments in the machine learning field are astonishing. Combined with digitalization (not just digitization, which means the process of converting something into digits), we are in the midst of the fourth industrial revolution.4,5)

A data-driven approach is suitable for the investigation of unknown things because it is neither biased nor restricted by existing knowledge. Among a host of analysis concepts and methods, we have recently devised a novel profile data analysis method, orthogonal linear separation analysis (OLSA), and have developed a new conceptual decomposition framework for understanding multiple effects of a drug in order to deepen our understanding of drugs.6) Here, we intentionally employed terminology profile data analysis instead of omics data analysis, although our methodology deals mainly with transcriptome data. In many cases, word profile data is regarded as a synonym of omics data, and the analysis methods dealing with omics data are regarded as omics data analysis. However, we consider that the concept behind the analysis methods employed in some newer studies, such as the OLSA and connectivity map (CMap), slightly differ from those of other well-used methods in omics analysis, such as pathway analysis and gene ontology (GO) analysis, which generate ambiguity in the word profile data analysis.7) In order to achieve high performance and specificity using these data analysis methods, it is indispensable to precisely understand the concept of the methodology of interest, and to design appropriate experiments for it.

Thus, our motivation in writing this review article is as follows: (1) to clarify any ambiguity in word profile data analysis by verbalization of the concepts behind the analysis method, and (2) to introduce profile data analysis in order to understand small compounds. This review begins with consideration of the meaning of the word profile data, followed by explanations of the framework of profile data analysis. The fourth and fifth sections introduce examples of profile data analysis methods employed in the understanding and utilization of the effects of drugs. In the sixth section, we discuss the methods for generating biological profile data, and finally, summarize the topics.

* To whom correspondence should be addressed. e-mail: tadahaya@mol.f.u-tokyo.ac.jp; kusuhara@mol.f.u-tokyo.ac.jp

© 2020 The Pharmaceutical Society of Japan
2. CLARIFICATION OF AMBIGUITY IN WORDS: PROFILE DATA ANALYSIS

First, we clarify what profile data is: profile data is multivariate data that describes an object. The definition of a word profile seems to be ambiguous, not only in the life sciences but also in general. Oxford Learner's Dictionaries, for instance, tells us the meaning of ‘profile’ is a description of somebody/something that gives useful information (https://www.oxfordlearnersdictionaries.com/); this does not give us clear understanding of the word. By contrast, the definition of ‘profiling’ offers relatively abundant information: the act of collecting useful information about somebody/something so that you can give a description of them or it. Based on these, we can understand that a profile is a description of an object composed of multiple data points, and profile data is this multivariate data. For example, gathered information about a criminal, such as height, sex, and hair color, and transcriptome data of liver dissected from a gene X knockout mouse, are multivariate data that describe the criminal and the mouse liver, respectively. It is an important property of profile data that it has not only values of each variate but also the relationship between them (intra-sample relationship); one variate alone is often of less meaning. It is not likely that tips about the height of a criminal alone greatly aids a criminal investigation, or the expression level of a gene deepens the understanding of the molecular mechanism working in the specimen, although it depends on the situation.

Considering the definition and the properties of profile data, profile data analysis is a method that deals with a data set composed of several profile data, employs both the intra- and inter-sample relationship (data structure, in other words,) and obtains knowledge about the specimen of interest (Fig. 1). Imagine that an existing drug A causes an increase of

---

**Biography**

Dr. Tadahaya Mizuno worked in molecular biology of transporter proteins and received the Ph.D. in pharmaceutical sciences from molecular pharmacokinetics laboratory, the University of Tokyo, Japan, in 2014. After graduation, he joined the laboratory as an assistant professor and keeps working currently. In addition to the above, he majors in bioinformatics. His current research interest is the potential of low molecular weight compounds such as drugs, particularly unidentified aspects of these. Interpretability is emphasized in his work, which comes from his background as a molecular biologist. By integrating molecular biology, pharmacokinetics, and bioinformatics, he tackles with understanding and utilization of the potential of low molecular weight compounds by unbiased approaches such as unsupervised matrix decomposition and exploratory factor analysis.
genes $a$ and $b$, and a decrease of gene $c$, while another drug B increases genes $a$ and $c$, but decreases $b$. In this situation, when a new drug X causes an increase in genes $a$ and $b$, and a decrease of gene $c$, in this case, we intuitively estimate $X$ is of similar effect to that of $A$. Notably, this estimation does not depend on what $a$, $b$, and $c$ are. Instead, it employs the intra-sample relationship between the variates $a$, $b$, and $c$ of each profile data, and the inter-sample relationship of the intra-sample relationship across the specimens. Therefore, one of the advantages of profile data analysis is its robustness against variate variation or absence. Although it is difficult to imagine the advantage in the above example, it turns out to be relatively easy to estimate the effects of drug X by using microarray data (approx. 10000 genes, in general cases) of cells treated with 30 drugs. In this situation, missing the data of 1000 genes would not affect the conclusion of which drug was similar to X, because the data structure would not be affected by those missing.

3. FRAMEWORK OF PROFILE DATA ANALYSIS

Profile data analysis is composed of three sequential processes: (1) data acquisition, (2) data analysis, and (3) interpretation of the output (Fig. 2). Extracting the biological information of a specimen without any bias would enable us to obtain profile data that could handle whole information, including unidentified relationships. The approach which employs omics data as profile data is one solution to this difficult request with regard to bias derived from an existing body of knowledge in biological knowledge, which restricts the scopes of analysis. For now, the outcomes of these methods are restricted to what we know so far, although they are incredibly powerful methods in understanding what happens in the specimen of interest. We have skipped detailed explanations of each method here because it is beyond the focus of this review, and excellent articles addressing these exist elsewhere.8–10)

On the other hand, there are profile data analysis methods that handle omics data but do not depend on existing bodies of biological knowledge. This type of profile data analysis is similar to that of the above well-known omics data analyses. The sole difference, with regard to existing knowledge independence, generates relatively large differences in profile properties. Due to the existing knowledge independence of this data analysis process, such profile data analysis has the advantage of potentially discovering novel findings. For instance, because it utilizes whole information, rather than existing knowledge only, these analysis outcomes could include unidentified biological phenomena, and even uncharacterized biological phenomena in the analysis process, both of which are suitable properties for novel findings. After analysis, interpretation of these analysis outcomes can be achieved by further biological knowledge, and unexplained outcomes are expected to produce novel findings.

How do profile data analyses, independent from biological knowledge, achieve these favorable properties? There are mainly two approaches. One utilizes mathematical hypotheses to interpret data structure. One such method is OLSA, which assumes that biological responses to small compounds can be explained by a linear combination of more basic responses to some extent.6) Network analysis used in many transcriptome data analyses employs mathematical hypotheses based on complex networks as well.11–14) Another approach employs
different data sets derived from different sources of a specimen of interest, including unidentified biological phenomena. One of the best examples of the former type is CMap, which employs the transcriptome data of cells treated with more than 1000 small compounds and compares them with profile data of interest.\footnote{7} However, to conduct this kind of profile data analysis, it is indispensable to prepare a profile data set that is appropriately acquired and processed, and satisfies the prerequisite of the analysis. Particularly, omics data is influenced by many undesirable variable factors such as differences in study centers, handlers, and instruments. These potential variants in batch effects should be taken into consideration, because they greatly affect the data structure of a profile data set.\footnote{15} To address this variability, there exist many normalization methods, such as quantile normalization, robust z normalization, and ranking conversion, combat, and Surrogate Variable Analysis (SVA).\footnote{16–18}

4. DECOMPOSITION OF MULTIPLE EFFECTS OF A DRUG: ORTHOGONAL LINEAR SEPARATION ANALYSIS

Profile data analysis can be a powerful tool in understanding a specimen of interest. We introduce one of the popular fields for these applications: the understanding and utilization of small compounds, including drugs, focusing on profile data analysis independent of biological knowledge. The basic concept of this application is that the biological responses of cells treated with a small compound reflect the profile of the chemical. Thus, the profile data of a chemical is the omics data of cells treated with the chemical. In this section, we introduce OLSA as one example of such applications for the understanding and utilization of drug effects (Fig. 3).

Recent success in drug repositioning indicates that small compounds, such as drugs, have unrecognized effects, even by the developers, and these effects are not single but multiple. However, we usually recognize only the main strong effect; the weak minor effects are often missed. Understanding such minor effects is expected to contribute to drug discovery, for example, avoiding toxicity and expanding the chemical space of a drug based on structure development, with a focus on effects. To this end, we invented a simple approach to decompose and understand the multiple effects of a chemical by applying unsupervised matrix decomposition to a profile data set, composed of the values reflecting chemical effects, assuming that the obtained mathematical structure preserves the original biological meanings.\footnote{6} OLSA does not require biological knowledge, since it is unsupervised. Here, unsupervised means analysis that does not utilize any pre-defined labels, whereas a supervised method requires such labels.

There are many kinds of unsupervised matrix decomposition methods.\footnote{19,20} Among these, we have selected the factor analysis principal component method because the outcomes of its analysis are relatively easy to interpret, due to its linear property approach. Because most biological phenomena are considered to be non-linear processes, in general, it is difficult to grasp where the outcomes of machine learning would be derived, which thus reduces the chance to utilize existing bodies of biological knowledge and human wisdom.\footnote{21} The efficiency of confirmation studies decreases without clear annotation by biological knowledge, even though the multiple effects of a target compound are decomposed. We have simplified conventional factor analysis by adding the mirror data set of the examined data set. Here, a mirror data set is a set of data

Fig. 3. Flow of Orthogonal Linear Separation Analysis

Graphic explanation of flow of OLSA. Profile data of a target compound is generated by omics analysis of cells treated with the compound of interest. Then, the data is decomposed with OLSA. Each decomposed effect is represented as a vector composed of multiple variates, such as genes in the case of transcriptome profile data. Coefficients of vectors in linear combination correspond to the strength of the decomposed effects. (Color figure can be accessed in the online version.)
points symmetrical to the original data set with regard to the origin. One concern in utilizing factor analysis is that the centroid in the novel coordinate space has no biological meaning, and varies among data sets, which means that the obtained factors in such a situation may not correspond to consistent biological meanings. However, mirror data enables us to mathematically approximate the novel coordinate space centroid to the origin of the original data space; this procedure assumes that antagonism exists when a biological response occurs, which biologically agrees with a situation in handling and developing drugs. Utilization of normalized data also gives the equivalent result mathematically, but the biological meaning of the procedure is difficult to determine. Note that OLSA does not employ outlier samples with regard to total strength of variates in preparation of the mirror data set, because biological reversibility cannot be assumed in such samples.

OLSA outcomes are a linear combination of the decomposed effects, each of which is a column vector of variables, such as genes. Coefficients of the vectors reflect the strength of each decomposed effect. For instance, in applying OLSA to a profile data set obtained from CMap, we were able to detect biologically consistent outcomes, such as a decomposed effect, which exhibits high positive scores in estrogens and high negative (opposite) scores in anti-estrogens. Based on outcomes, the autophagy inducibility of some small compounds was experimentally confirmed. These results indicate that OLSA has the potential to uncover unidentified aspects of small compounds. We are currently working on the detection of latent toxicity of drugs, and understanding multiple effects of a natural product (data not shown).

5. OTHER APPLICATIONS OF PROFILE DATA ANALYSIS IN THE UNDERSTANDING AND UTILIZATION OF DRUG EFFECTS

In this section, we introduce other applications for the understanding and utilization of drug effects. The first useful application is understanding a target compound with unknown effects simply by comparing its profile data with compounds with known effects. This is the simplest application, but powerful nevertheless. Imagine we have a profile data set composed of well-characterized compounds A, B, and C, and unknown compound X, which shows a similar profile to that of A. Compound X is estimated to have similar effects to those annotated with compound A. Because there are several available similarity measures, such as cosine distance, Euclidian distance, and Mahalanobis’ distance, relatively similar outcomes can be obtained in many cases, and calculation itself is easy.

This approach works particularly well in uncovering the effects of natural compounds because it can estimate the effects without speculating a possible effect, after which we can establish an evaluation system, and then run experiments, one by one.

The second application takes a more data-driven approach, employing the reversed data of the profile data of a specimen of interest, which assumes that a compound which reverses the profile data has an antagonizing effect against the specimen’s phenotype. In fact, using CMap, the profile data of estrogens has been shown to be similar to the reversed profile data of estrogen antagonists. One of the sophisticated examples of reversed data profiling is the reprogramming of docetaxel-resistant prostate cancer. Kosaka et al. acquired transcriptome data of both docetaxel-resistant prostate cancer model cells and normal prostate cancer cells, and searched for drugs that could reverse the difference between them. Amazingly, the combined treatment of docetaxel and one of the hit drugs, Rivabirin, actually decreased the tumor volume of mice injected with docetaxel-resistant cells. Notably, this reprogramming treatment has worked well in clinical trials.

The above work implies that a chemical which reverses the profile data of a disease has potential to become a medication to treat the disease. The important thing, in the words of Dr. Horimoto (the corresponding author of the reprogramming study above), is connecting drugs and diseases with data using the 1:many:1 relationship (Fig. 4). Data is thus employed as a common language between a compound and a disease, which are two completely different things. This kind of application is a very data-driven, innovative approach. To the best of our knowledge, Kosaka’s study is the only one to have entered clinical trials, though many studies have reported success in vivo. There are many reasons for difficulties in in vivo extrapolation, such as pharmacokinetics. Combined with progress in these areas, the reversing profile data approach will become an increasingly powerful one.

6. METHODS FOR COMPREHENSIVE DIGITIZATION OF A BIOLOGICAL SPECIMEN

Profile data subjected to the analysis methods described above should be acquired without usage of the existing body of biological knowledge. To achieve this, the important thing
is comprehensiveness, because such profile data analysis assumes that comprehensively acquired data of a specimen is numeric information that reflects the biological information of the specimen. Here, we introduce such comprehensive digitization of a biological specimen, from the point of view of profile data analysis.

Today, transcriptome data is the best choice for converting profile data of a small compound into data which can be used in research. First, in cases of the most employed transcriptome profiles, such as RNA-sequence and microarray, approximately 10000 variates, or about half of all human genes, can be obtained from any specimen; this number overwhelms those of other methods, such as metabolome and proteome analysis. Moreover, it is possible to utilize biological knowledge of the variables in the interpretation process of analyzing outcomes, since they clearly correspond to genes. It is also noteworthy that databases of transcriptomes are well established: they store amazing amounts of data (https://www.ncbi.nlm.nih.gov/gds, https://www.ebi.ac.uk/arrayexpress/). CMap, the pioneering process of profile data analysis of small compounds, employs transcriptome data obtained via a microarray platform. Recently, the next generation project of CMap was launched, termed the Library of Integrated Network-based Cellular Signatures (LINCS).30,31 LINCS employs transcriptome profile data, although with added ingenuity to achieve high-throughput data acquisition. LINCS narrows target genes by data-driven analysis of 12063 transcriptome profile data points, so that the genes ultimately selected explain sufficient biological information about a specimen. Currently, LINCS stores more than 1 million profile data of small compounds (http://www.ilincs.org/ilincs/).

On the other hand, to best of our knowledge, other well-known omics data, such as proteome and metabolome data, have not yet employed profile data describing small compounds. Some reasons for this are that the number of variates these can handle is relatively small; therefore these databases remain to be comprehensively established compared with those of transcriptome. However, the potential use for proteome and metabolome data as profile data is likely high. Differences between the layers are expected to extract different profiles of a specimen, and to increase the accuracy of its description. Considering central dogma, proteome and metabolome profile data may describe phenotypes of a specimen more sharply than transcriptome profiles do. Notably, the SWATH-MS method (sequential window acquisition of all theoretical fragment ion spectra mass spectrometry), a novel LC-MS/MS method, greatly improves the accuracy and number of observable variates, compared with those of a data-dependent acquisition method, which has been mainly employed by omics analysis using LC-MS/MS.32,33) This potential improvement in comprehensive data collection is good news for profile data analysis using proteome and metabolome data.

There exist other acquisition methods of profile data of small compounds. In 2007, Young et al. reported a profile data acquisition method based on morphological changes in cells treated with small compounds.34) Recently, Carpenter and colleagues, who belong to the Broad Institute, have invented a novel method using a similar concept, by employing cell-image-based changes detected with a high content analyzer.35,36) In another interesting example, Muroi et al. established a 2-dimensional electrophoresis (2DE)-based platform for obtaining proteome profile data of small compounds, and succeeded in determining the mode of action of a natural product.24) Recently, we have also established a 2DE-based platform that overcomes some of the classic weaknesses of 2DE, such as low handling due to a large size gel, and manual handling of image analysis.37) Although 2DE-based proteome profile data has a clear weak point, in that the annotation of each variable is difficult to determine, the attractive point is that it can detect proteome profile data composed of intact proteins without any processing such as trypsinization. Thus, with 2DE-based proteome profiling, the concept of describing a specimen with comprehensive data is not restricted to so-called omics analysis data. It may be interesting to expand the applications of this concept to clinical big data, such as spontaneous reporting system data of adverse events and receipt data.

7. CONCLUDING REMARKS

Hypothesis construction and validation creates research. However, in terms of drug discovery, these classic methods are expensive in terms of both time and cost.39) It is obvious that the existing body of biological knowledge is useful and invaluable human wisdom. Yet, what we currently know about biological phenomena is probably the tip of the iceberg. Therefore, hypotheses dependent on our existing body of biological knowledge may explain only part of the whole realm of biological phenomena. Drug discovery would remain quite arduous if we tackled it with only a hypothesis-driven approach. In drug discovery, we need to understand both the target (a disease) and its perturbation (a drug). There are currently approaches independent of our current body of biological knowledge that allow us to acquire information from both sides, disease and drug, such as omics analysis of clinical specimens, and high-throughput screenings using a chemical library, for the former and the latter, respectively.39,40) However, in current research practice, after data acquisition, each side is connected using separate biological knowledge, then the two sides are bridged, in many cases, by this biological knowledge. Compared with this, profile data analysis, such as CMap, directly connects the disease side and the drug side by utilizing data as a common language, and is thus extremely powerful. This capability may be especially effective for drug discovery to treat diseases whose mechanisms are complicated, such as atherosclerosis and diabetes. With regard to the above, low molecular weight drugs are still attractive, since they have multiple effects, although the modality tends to shift to high and medium molecular weight drugs.31) Unrecognized aspects of small compounds, even by developers, are expected to explain the characters of diseases that currently cannot be discerned by the existing body of biological knowledge, and as such, have promise as attractive tools in the life sciences field. Profile data analysis, such as OLSA, could support these kinds of applications. The important point is that each method has strengths and weaknesses. Therefore, appropriate data acquisition, data analysis, and interpretation must be done, which calls for a clear recognition of the differences among the concepts behind these analysis methods.

Along with the fourth industrial revolution, the style of life science research is changing. We expect that a combination of the conventional hypothesis-driven approach and the data-driven approach, such as profile data analysis, will synergisti-
cally promote drug discovery and the life sciences in general.

**Acknowledgments** This study was financially supported by a Grant-in-Aid for Challenging Exploratory Research (17K19478) from the Japan Society for the Promotion of Science, and by a Grant-in-Aid from The Mochida Memorial Foundation for Medical and Pharmaceutical Research. We are grateful for helpful discussions with Dr. Katsuhisa Horimoto at the Molecular Profiling Research Center for Drug Discovery (molprof), at the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo.

**Conflict of Interest** The authors declare no conflict of interest.

**REFERENCES**

1. Gyllensten H, Hakkarainen KM, Hägg S, Carlsten A, Petzold M, Rehnberg C, Jönsson AK. Economic impact of adverse drug events — a retrospective population-based cohort study of 4970 adults. *PLoS ONE*, 9, e92061 (2014).

2. Lipton SA. Failures and successes of NMDA receptor antagonists: molecular basis for the use of open-channel blockers like memantine in the treatment of acute and chronic neuropsychologic insults. *NeuroRx*, 1, 101–110 (2004).

3. Hayashi H, Naoi S, Hirose Y, Matsuzyaka Y, Tanikawa K, Igarashi K, Nagasaka H, Kage M, Inui A, Kusuhara H. Successful treatment with 4-phenylbutyrate in a patient with benign recurrent intrahepatic cholestasis type 2 refractory to biliary drainage and bilirubin absorption. *Hepatol. Res.*, 46, 192–200 (2016).

4. Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. *Cell.*, 177, 58–69 (2019).

5. National Academies of Sciences, Engineering and MP and GAG-U-IRR. The Fourth Industrial Revolution: Proceedings of a Workshop—in Brief. National Academies Press, Washington, D.C. (2017).

6. Mizuno T, Kinoshita S, Itō T, Maedera S, Kusuhara H. Development of orthogonal linear separation analysis (OLSA) to decompose drug effects into basic components. *Sci. Rep.*, 9, 1824 (2019).

7. Lamb J, Crawford BE, Peck D, Modell JD, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reimand J, Schaub PM, Jeffery BM, Datta S, Barrette TR, Gerstein M, Wang Z, Coifman RR, Stawicki P, Long CH, Botstein D, Brown PO. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463, 318–325 (2010).

8. Mansurilli N, Astefanoaei D, Di Matteo T. Relation between financial market structure and the real economy: comparison between clustering methods. *PLoS ONE*, 10, e0116201 (2015).

9. Nyaasang V, Rodland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17, 29–39 (2016).

10. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8, 118–127 (2007).

11. Yeek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, 3, 1724–1735 (2007).

12. Haghverdi L, Lun ATL, Morgan MD, Mariotti JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 36, 421–427 (2018).

13. Liu C, Su J, Yang F, Wei K, Ma J, Zhou X. Compound signature detection on LINCS L1000 big data. *Mol. Biosyst.*, 11, 714–722 (2015).

14. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Netw.*, 13, 411–430 (2000).

15. Faranajpe K, Schinkel M, Nanayakkara P. Short keynote paper: mainstreaming personalized healthcare-transforming healthcare through new era of artificial intelligence. *IEEE J. Biomed. Health Inform.*, 24, 1860–1863 (2020).

16. Tan P-N, Kamar M, Srivastava J. Selecting the right objective measure for association analysis. *Inf. Syst.*, 29, 293–313 (2004).

17. Leena P, Meyer P, Vaillant B, Lalles E. On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. *Eur. J. Oper. Res.*, 184, 610–626 (2008).

18. Muroi M, Kazami S, Noda K, Kondo H, Takayama H, Kawatani M, Usui T, Osada H. Application of proteomic profiling based on 2D-DIGE for classification of compounds according to the mechanism of action. *Chem. Biol.*, 17, 460–470 (2010).

19. Kosaka T, Nagamatsu G, Saito S, Oya M, Suda T, Horimoto K. Identification of drug candidate against prostate cancer from the aspect of somatic cell reprogramming. *Cancer Sci.*, 104, 1017–1026 (2013).

20. Schenck EL. Clinical outcomes of patients on check point inhibitor therapy who receive routine vaccinations. *J. Clin. Oncol.*, 35 (15 suppl), e14597 (2017).

21. Kunkel SD, Suneja M, Ebert SM, Bongers KS, Fox DK, Malmborg SE, Alipour F, Shields RK, Adams CM. mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell Metab.*, 13, 627–638 (2011).

22. Brum AM, Van De Peppel J, Van Der Leije CS, Schreuder-Holtrop M, Eijken M, Van Der Eerden BJ, Van Leeuwen JPM. Connectivity map-based discovery of parbendazole reveals targetable human osteogenic pathway. *Proc. Natl. Acad. Sci. U.S.A.*, 112, 12711–12716 (2015).

23. Manzotti G, Parenti S, Ferrari-Amorotti G, Soliera AR, Cattelan S, Montanari M, Cavalli D, Ertel A, Grande A, Calabretta B, Monocyte-macrophage differentiation of acute myeloid leukemia cell lines by small molecules identified through interrogation of the connectivity map database. *Cell Cycle*, 14, 2578–2589 (2015).

24. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 100000 Profiles. *Cell*, 171, 1437–1452.e17 (2017).

25. Statthias V, Turner J, Koleti A, Vidovic D, Cooper D, Fazel-Najafabadi M, Pilarczyk M, Terryn R, Chung C, Umeno A, Clarke DB, Lachmann A, Evangelista JE, Ma'ayan A, Medvedovic M, Schürer SC. LINCS Data Portal 2.0: next generation access point for perturbation–response signatures. *Nucleic Acids Res.*, 48 (D1), D431–D439 (2020).

26. Schubert OT, Gillet LC, Collins BC, Navarro P, Rosenberger G, Wolski WE, lam H, Amodei D, Mallick P, MacLean B, Aebersold R.
R. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.*, **10**, 426–441 (2015).

33) Guo T, Kouvonen P, Koh CC, Gillet LC, Wolski WE, Röst HL, Rosenberger G, Collins BC, Blum LC, Gillessen S, Joerger M, Jochum W, Aebersold R. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat. Med.*, **21**, 407–413 (2015).

34) Young DW, Bender A, Hoyt J, McWhinnie E, Chirn GW, Tao CY, Tallarico JA, Labow M, Jenkins JL, Mitchison TJ, Feng Y. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, **4**, 59–68 (2008).

35) Bray MA, Singh S, Han H, Davis CT, Borgeson B, Hartland C, Kost-Alimova M, Gustafsdottir SM, Gibson CC, Carpenter AE. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.*, **11**, 1757–1774 (2016).

36) States C, Gustafsdottir SM, Ljosa V, Sokolnicki KL, Wilson JA, Walpita D, Kemp MM, Seiler KP, Carrel HA, Golu TR, Schreiber SL, Clemons PA, Carpenter AE, Shamji AF. Multiplex cytological profiling assay to measure diverse cellular states. *PLoS ONE*, **8**, 1–7 (2013).

37) Kinoshita S, Mizuno T, Hori M, Kohno M, Kusuhara H. Development of a novel platform of proteome profiling based on an easy-to-handle and informative 2D-DIGE system. *Biol. Pharm. Bull.*, **42**, 2069–2075 (2019).

38) Dima JS, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.*, **47**, 20–33 (2016).

39) Benjamin D, Colonbi M, Hindupur SK, Betz C, Lane HA, El-Shemerly MYM, Lu M, Quagliata L, Terracciano L, Moes S, Sharpe T, Wodnar-Filipowicz A, Moroni C, Hall MN. Syrosingopine sensitizes cancer cells to killing by metformin. *Sci. Adv.*, **2**, e1601756 (2016).

40) Pournoor E, Mousavian Z, Dalini AN, Masoudi-Nejad A. Identification of key components in colon adenocarcinoma using transcriptome to interactome multilayer framework. *Sci. Rep.*, **10**, 4991 (2020).

41) G. de la Torre B, Albericio F. The pharmaceutical industry in 2018. An analysis of FDA drug approvals from the perspective of molecules. *Molecules*, **24**, 809 (2019).