Computational Approaches and Pharmacogenomics Data Resources for Drug Repositioning

Author:
Fuhai Li

Authors Note:

Department of Biomedical Informatics,
The Ohio State University,
 Columbus, OH, USA
Email:
Fuhai.Li@osumc.edu

Abstract
Large-scale data sets of cancer patients have been being generated due to the significantly reduced cost of sequencing full genome of individual patients using the Next Generation Sequencing (NGS) technology. Comprehensive genomics data analysis revealed the diverse dysfunctional biomarkers of individual cancer patients, which are believed to be responsible for heterogeneous drug response. Thus precision medicine is becoming popular that aims to find the optimal treatments for individual patients based on their genomics profiling data. However, it is challenging to interpret the complicated and distinct genome mutation and variation patterns, and associate them to optimal treatments. Though a set of approaches and data resources have been reported to reposition FDA approved drugs and investigational drugs for specific diseases, novel and sophisticated computational approaches are needed urgently to reposition drugs for cancer subtypes or individual patients. In this study, some widely used computational approaches and pharmacogenomics data resources for repositioning optimal drugs are introduced and discussed, which aims to provide a general overview of the genomic data-driven drug repositioning, and help readers understand the topic conveniently.

Key words
Precision medicine, Personalized medicine, Drug repositioning, Drug combination, Genetic medicine

Computational Approaches and Pharmacogenomics Data Resources for Drug Repositioning

1. The need and feasibility of data-driven drug repositioning

There is no routine cancer. Due to the advance of the next generation sequencing (NGS) technology (1), it becomes possible to sequence a human genome with around US$1,000 (2). Consequently, large-scale datasets of genomics of cancer patients have been being generated rapidly. For example, in The Cancer Genome Atlas (TCGA), comprehensive genomics data sets, e.g., mutation, methylation, copy number aberration, gene expression, and protein expression of 14551 cases (samples) from 38 types of cancer (by March 21, 2017) are publically available at National Cancer Institute (NCI's) Genomic Data Commons (GDC) website (GDC URL: https://gdc.cancer.gov/). Another convenient way to access TCGA data and other cancer genomics data sets is using the cBioPortal (3) (cBioPortal URL: http://www.cbioportal.org/). The genomics data sets can be download using the R (a widely used free software for statistical data analysis, and for the bioinformatics data analysis) Package 'cgdsr', which provides a set of R functions of data accessing/downloading from the Cancer Genomics Data Server (CGDS), hosted by the Computational Biology Center at Memorial-Sloan-Kettering Cancer Center (MSKCC). Systematic genomics data analyses based on the TCGA data sets have revealed diverse cancer subtypes bearing distinct dysfunctional biomarkers (4-6), which are believed to be responsible for heterogeneous drug response. Another two projects, i.e., the Cancer Cell Line Encyclopedia (CCLE) (7) and Genomics of Drug Sensitivity in Cancer (GDSC) (8) were reported in 2012 to show the association of genomics biomarkers and drug response of about 1000 cancer cell lines. The drug response data and genomics profiles of these cancer cell lines are publically available (CCLE URL: https://portals.broadinstitute.org/ccle/home and GDSC URL: http://www.cancerrxgene.org/). Therefore,

personalized treatment (precision medicine) or health care should be designed or tailored for individual patients based on unique biomarkers embedded in the personal genomics data as well as the lifestyle and environment factors (9). In early 2015, the precision medicine initiative (PMI) was launched by president Barack Obama (10). One of the short-term goals of PMI is to expand the precision medicine in cancer research to find new and more effective treatments for various kinds of cancer patients based on the increased knowledge of the heterogeneity and landscape of cancer genetics. Since it is challenging to select effective drugs from thousands of FDA approved drugs and investigational drugs with for individual patients and cancer subtypes bearing distinct and complicated genome mutations and variations, genomics data-driven computational approaches are needed urgently to associate personal genomics data with optimal drugs (11) or drug combinations (12).

For easy understanding, in this study, drug repositioning that discovers new uses or new indications of existing drugs (FDA approved and investigational drugs) (13) will be used to represent discovering optimal drugs for individual patients or cancer subtypes based on personal genomics profiles. Though a set of computational approaches and data resources have been reported for drug repositioning (11). However, it remains an open problem to find (reposition) the new and more effective drugs for individual cancer patients based on the personal genomics data. There are two major challenges. First, there is no effective drug that can directly target on the dysfunctional or mutated genes. Secondly, cancer is a complex disease that is not controlled by single gene defects, but regulated by a group of coordinated genes (14). The aim of this study is to provide an overview of the genomics data-driven drug repositioning. Some widely used computational approaches and pharmacogenomics data resources for drug

repositioning are summarized and discussed in the following sections.

2. Reverse gene expression signature based drug prediction

In 2006, the "Connectivity Map" data was reported in Science (11) that proposed to connect drugs, genes and diseases via the 'similarity' among gene expression signatures (i.e., a set of gene selected based on fold change or p-value) in a set of tumor cell lines caused by perturbations of chemical (drugs) and genetic (genes) reagents. The gene expression signatures of diseases can be obtained by comparing the gene expression between disease tissues and normal tissues. To estimated the similarity score between two given gene expression signatures, the Gene Set Enrichment Analysis (GSEA) (15) approach is widely used. Figure 1 illustrates the reverse GSEA score based drug repositioning in CMAP. Connectivity Map (CMAP) is the most widely used and successful approach and data resource for drug repositioning (11), which has been widely used to reposition drugs for various types of cancers (16, 17). The CMAP data resources and tools for repositioning drugs using CMAP are accessible at URL: https://clue.io/. With the disease gene signature, users can query (reposition) drugs directly from the above CMAP website. More CMAP data and other perturbation data sets can be found at The Library of Network-Based Cellular Signatures (LINCS) Program URL: http://lincsportal.ccs.miami.edu/dcic-portal/.
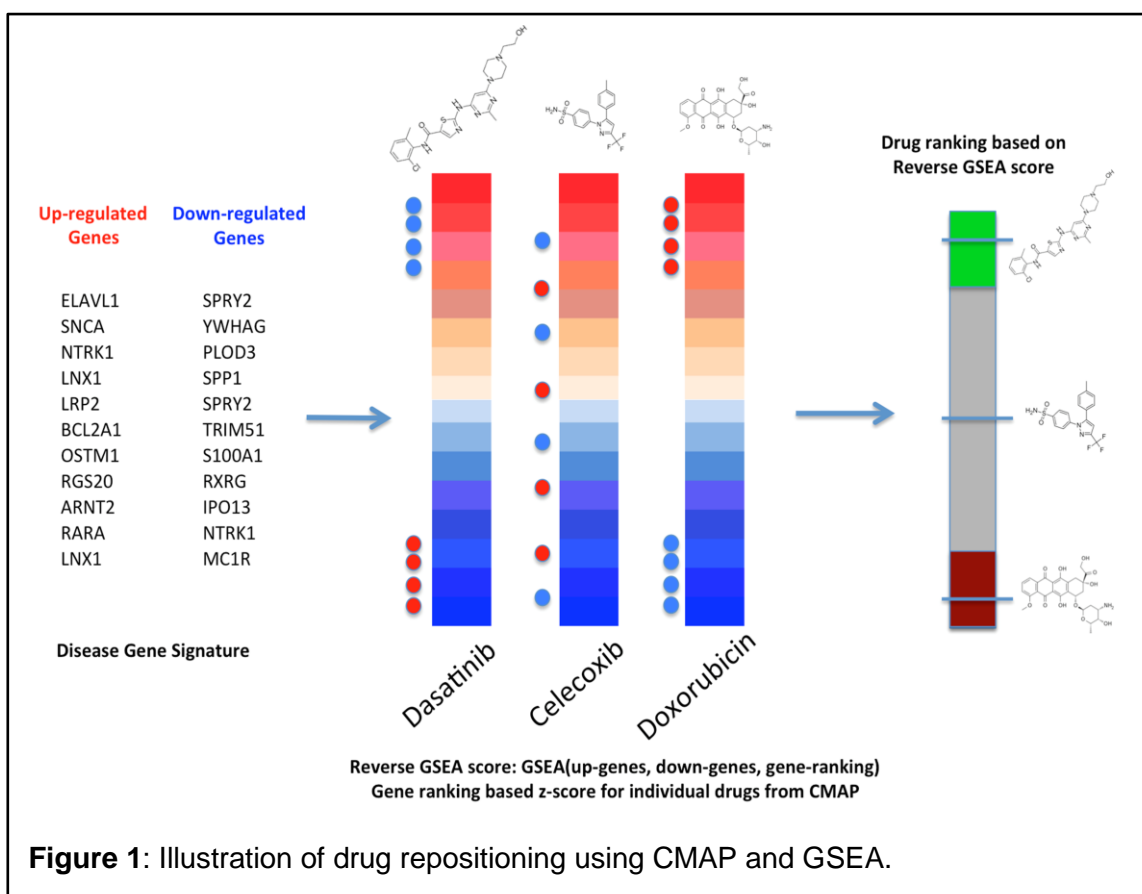
Figure 1: Illustration of drug repositioning using CMAP and GSEA.

Many improved gene expression signature based approaches have been reported based on the GSEA approach in CMAP. Since the gene expression signature based drug repositioning results closely depend on the selection of gene signatures, the random disease signatures were generated to statistically evaluate the performance of a drug on the selected disease signature (compared with the scores on the random disease signatures) (18). Multiple gene signatures might be able to generate more reliable drug repositioning results. For example, integrative meta-analysis was proposed to integrate multiple drug ranking lists (obtained from multiple gene signature
.

3. Network-based drug repositioning

As aforementioned that cancer is a complex disease that is not controlled by single gene defects, but regulated by a group of coordinated genes (14). Thus network medicine or network-based approaches should be used to reposition drugs (14, 22). In (22), network-based proximity was used to study the distance between drug targets and disease genes, and concluded that the effective drugs' targets are localized in the network neighborhood of the disease genes. This observation indicates a new way of drug repositioning by prioritize drugs whose targets are nearby the disease genes or disease signaling network (22, 23).

The disease associated genes can be obtained from the OMIM database (24) and the GWAS catalog (25). Also the disease associated genes can be conveniently obtained from DisGeNET (a database of gene-disease associations) (26) (DisGeNET URL: http://www.disgenet.org/web/DisGeNET/menu). On the other hand, there are a set of protein-protein interaction and signaling pathway database (27) can be used as the background disease signaling network, e.g., BioGRID (28) (URL: http://thebiogrid.org/download.php), STRING (29) (URL: http://string-db.org/) and KEGG (30) (URL: http://www.genome.jp/kegg/pathway.html).

inputs) in (19). Also, the gene expression signatures can be combined with drug targets as well as chemical similarity of drugs to improve the drug repositioning performance (20). To learn more about drug repositioning, the NCI dream challenges (URL: http://dreamchallenges.org/project-list/closed/) also provided a few data sets of predicting drug response as well as drug combination (cocktails) response (21). Sage Synapse (URL: https://www.synapse.org/) hosts the data sets and provides computational resources for the convenient data access, analysis and solution sharing with different research groups.

Some computational approaches have been reported to construct the disease specific signaling networks by integrating the gene expression and disease associated genes on the background signaling network (31-35). In the following, as an example, a Glioblastoma (GBM) disease signaling network was constructed.
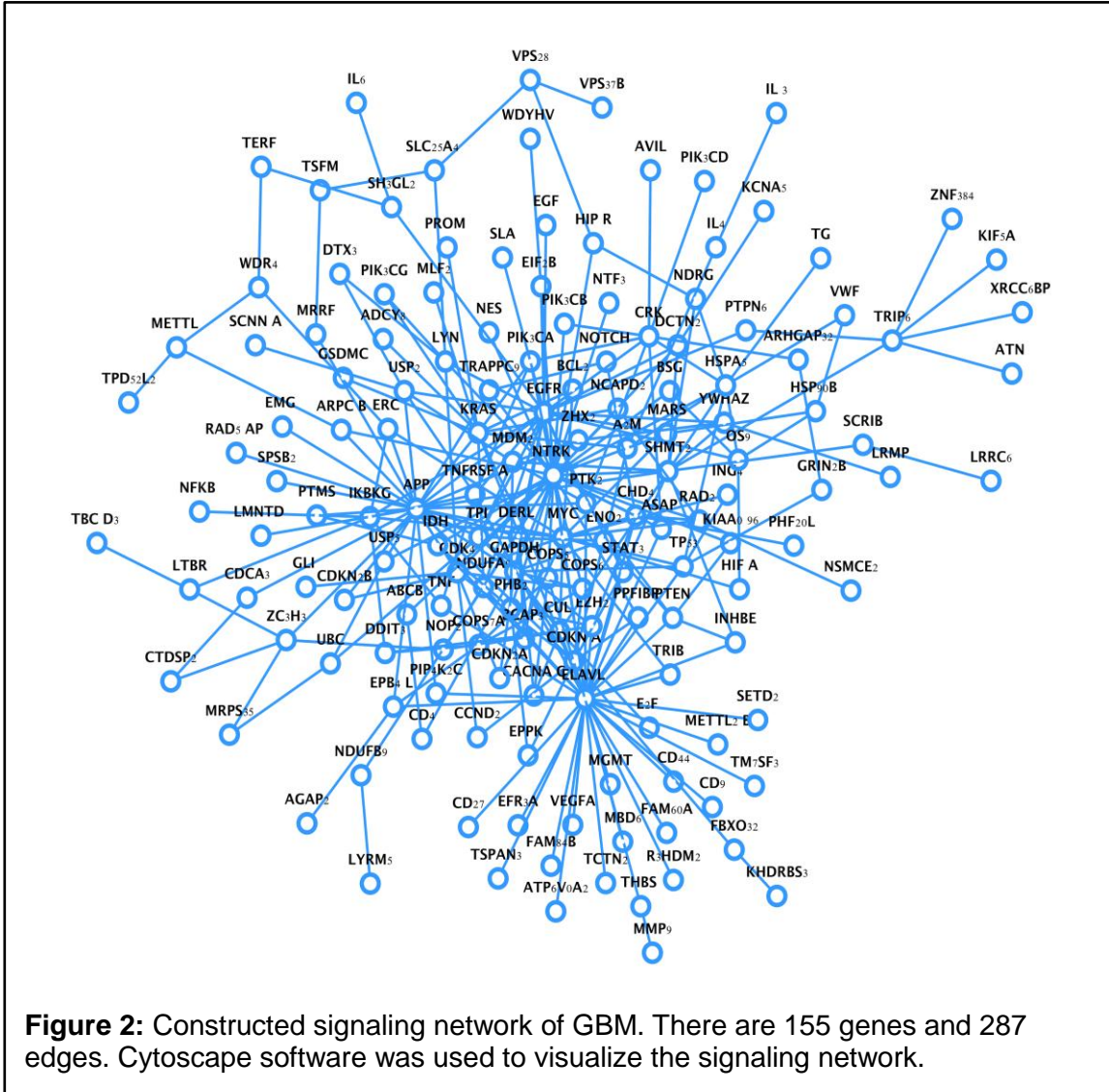
The gene expression data of GBM proneural subtype (the data were extracted from TCGA with IDH1 mutation 14 samples) (36) were obtained by using the R Package 'cgdsr'. In addition, the top 30 reported GBM associated genes were obtained from DisGeNET database (26). The protein-protein interaction data was obtained from the BioGRID database (28) as the background signaling network, and the self-interaction edges were removed. Then the belief propagation algorithm (33) was employed to construct the disease signaling network by modeling the signaling flow starting from the specified disease "root" genes and linking them to up-regulated genes on the BioGRID background signaling network. The top 30 ranked GBM-associated disease genes from the DisGeNET database were used as root nodes. Mathematically, Given the BioGRID background network, $G = (V, E)$, the sub-network, $G' = (V', E')$, is constructed to minimize the cost function:

$$\min_{E'\subseteq E, V'\subseteq V} \sum_{e\in E'} c_e - \lambda \sum_{i\in V'} b_i$$

where c_e (cost of edge) is 1.0, $\lambda = 0.1$, and b_i is the evidence score combining gene expression fold change and p-value as follows: $b_i = Z_i$, (z_i is the z-score of the i-th gene).

Figure 2 shows the constructed signaling network of the GBM subtype. Cytoscape (a widely used network visualization software with a large number of plug-ins for network-based analysis) was used to visualize the constructed signaling network (Cytoscape URL: http://www.cytoscape.org/).

Figure 2: Constructed signaling network of GBM. There are 155 genes and 287 edges. Cytoscape software was used to visualize the signaling network.

To reposition drugs, drugs' targets will be firstly mapped onto the back-ground network, and then calculate the drug targets and disease genes distance as reported in (22); or map the drug targets on the constructed signaling network, and then drugs will be prioritized (12, 34) using the betweenness, closeness and page-rank centrality metrics (37, 38). Drug target information can be obtained from DrugBank

(39) (URL: https://www.drugbank.ca/) and STITCH (40) (URL: http://stitch.embl.de/) database and other resources (41). **Figure 3** shows the targets of Dasatinib obtained from STITCH. In addition to single drug repositioning, synergistic drug combinations can also be repositioned computationally using the network-based approach, e.g., DrugComboRanker and GenSynNET (37, 38), as well as other approaches, e.g., Combinatorial Drug Assembler (CDA) (42, 43), DrugPairSeeker (DPS) (44).

**3. Clinical interpretation of genetic mutations**

With the comprehensive data analysis of the molecular profiling of a large number of cancer patients, the genetic heterogeneity and mutational landscape of various types of cancer are being revealed. These diverse genetic mutations were reported to be
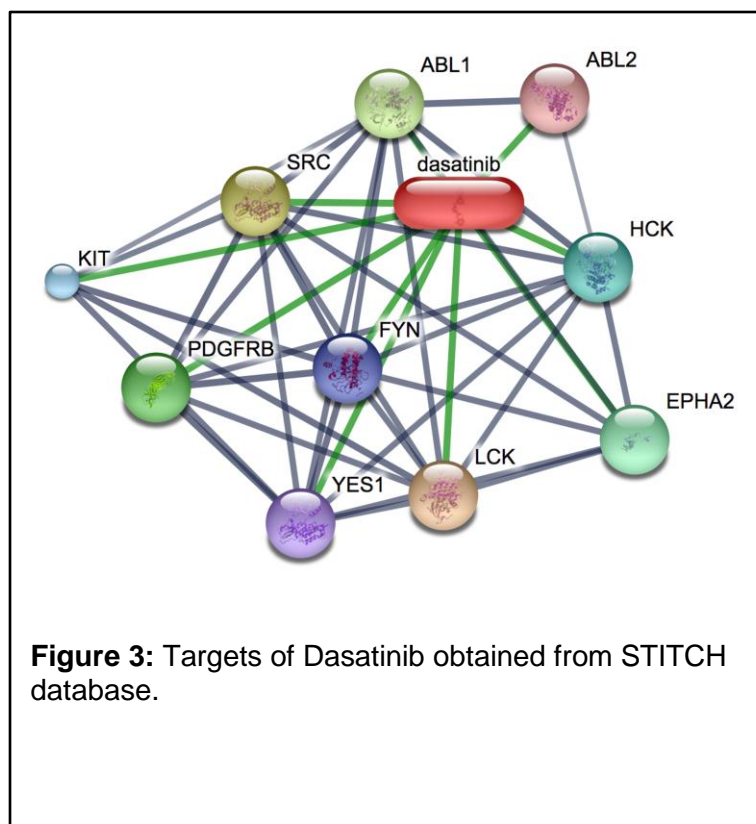
Figure 3: Targets of Dasatinib obtained from STITCH database.

associated with heterogeneous drug responses of individual patients in clinical. For example, the BRAF inactivating mutation was reported as a biomarker of Dasatinib sensitivity response in non-small cell lung cancer (45). There are a number of data resources available to facilitate the clinical drug selection based on patients' genetic mutations. For example, ClinVar (46, 47) (URL: https://www.ncbi.nlm.nih.gov/clinvar/) at the National Center for Biotechnology Information (NCBI). ClinGen (The Clinical Genome Resource.) (48) (URL: https://www.clinicalgenome.org/about/). MD Anderson Cancer Center's Personalized

Cancer Therapy Knowledgebase (URL: https://pct.mdanderson.org/) and Precision Medicine Knowledgebase (URL: https://pmkb.weill.cornell.edu/). In 2017, the open-source CIViC (clinical interpretation of variants in cancer) (49) (URL: http://civicdb.org/) was released for public access. **Figure 4** shows the 'NRAS' mutation query results from CIViC database. As can be seen, there are a few drugs have been reported to be associated to the NRAS mutation, but the Trust Rating (TR) and Evidence Level (EL) of some drugs are very low indicating more evidence needed to support the clinical use.

Figure 4: Query results of 'NRAS mutation' from CIViC database.
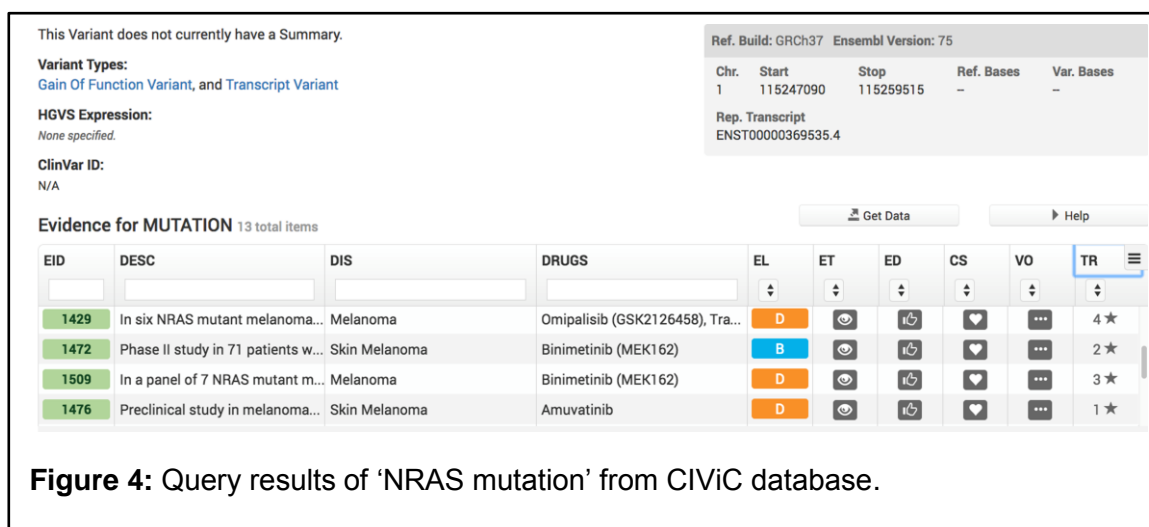
These data resources have been used to provide clinical evidence to guide the precision personalized treatment of cancer patients. For example, Foundation Medicine (URL: https://www.foundationmedicine.com/), which is a precision medicine company, has been providing precision medicine solutions based on the entire coding sequence of 315 cancer-related genes and introns from 28 genes often rearranged or altered for solid

tumor cancers. However, the landscape and heterogeneity of cancer genetic mutations are complicated. Multiple distinct genetic mutations often appeared in individual patient's genome profiling. Thus it remains a challenging and open problem to associate precision medicine to individual patients base on the personal genome profiles. Therefore, sophisticated computational approaches are needed urgently to identify the accurate drugs precisely for individual

patients. The known clinical evidence can be used as the valuable training data to evaluate and improve the computational approaches.

4. Summaries and conclusions

The era of precision medicine especially in cancer has been coming due to the advance of the next generation sequencing technology. Many studies have been reported to discover the diverse and complicated driver biomarkers of various cancer subtypes and individual patients, which provide the basis for precision medicine or personalized medicine. However, precision medicine remains an open problem, as it is challenging to select the optimal drugs and drug combinations for individual patients with diverse and complicated genome mutation and variation patterns. Novel and sophisticated computational approaches will play important roles in precision medicine, and are needed urgently to accurately associate optimal drugs and drug combinations to individual patients based on their unique genome mutation and variation patterns. In this study, some widely used computational approaches and pharmacogenomics data resources for drug repositioning are reviewed and discussed to provide an overview of this field and help readers understand the topic of computational drug repositioning and precision medicine conveniently.

Reference

1. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586-97.

2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51.

3. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Science signaling. 2013;6(269):pl1-pl.

4. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.

5. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519-25.

6. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609-15.

7. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7.

8. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483(7391):570-5.

9. Hodson R. Precision medicine. Nature. 2016;537(7619):S49.

10. Collins FS, Varmus H. A New Initiative on Precision Medicine. New England Journal of Medicine. 2015;372(9):793-5.

11. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929-35.

12. Huang L, Li F, Sheng J, Xia X, Ma J, Zhan M, et al. DrugComboRanker: drug combination discovery based on target network analysis. Bioinformatics. 2014;30(12):i228-i36.

13. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3(8):673-83.

14. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56-68.

15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.

16. Jahchan NS, Dudley JT, Mazur PK, Flores N, Yang D, Palmerton A, et al. A Drug Repositioning Approach Identifies Tricyclic Antidepressants as Inhibitors of Small Cell Lung Cancer and Other Neuroendocrine Tumors. Cancer Discovery. 2013.

17. Paik H, Chung A-Y, Park H-C, Park RW, Suk K, Kim J, et al. Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records. Sci Rep. 2014;5.

18. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. Science Translational Medicine. 2011;3(96):96ra77-96ra77.

19. Fortney K, Griesman J, Kotlyar M, Pastrello C, Angeli M, Sound-Tsao M, et al. Prioritizing Therapeutics for Lung Cancer: An Integrative Meta-analysis of Cancer Gene Signatures and Chemogenomic Data. PLOS Computational Biology. 2015;11(3):e1004068.

20. Lee H, Kang S, Kim W. Drug Repositioning for Cancer Therapy Based on Large-Scale Drug-Induced Transcriptional Signatures. PLOS ONE. 2016;11(3):e0150460.

21. Bansal M, Yang J, Karan C, Menden MP, Costello JC, Tang H, et al. A community computational challenge to predict the activity of pairs of compounds. Nat Biotech. 2014;32(12):1213-22.

22. Guney E, Menche J, Vidal M, Barabasi AL. Network-based in silico drug efficacy screening. Nat Commun. 2016;7:10331.

23. Haoyang Wu EM, Denethi Wijegunawardana, Kelly Regan, Philip Payne, Fuhai Li. MD-Miner: A network-based approach for personalized drug prediction. ICIBM 2016 (to appear in Journal of BMC Systems Biology). 2016.

24. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-7.

25. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014;42(Database issue):D1001-D6.

26. Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015;2015:bav028.

27. Klingström T, Plewczynski D. Protein–protein interaction and pathway databases, a graphical review. Briefings in Bioinformatics. 2011;12(6):702-13.

28. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(Database issue):D535-9.

29. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):D447-52.

30. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28(1):27-30.

31. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559.

32. Stuart JM, Segal E, Koller D, Kim SK. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science. 2003;302(5643):249-55.

33. Bailly-Bechet M, Borgs C, Braunstein A, Chayes J, Dagkessamanskaia A, Francois JM, et al. Finding undetected protein associations in cell signaling by belief propagation. Proc Natl Acad Sci U S A. 2011;108(2):882-7.

34. Regan K, Payne P, Li F. Integrative network and transcriptomics-based approach predicts genotype-specific drug combinations for melanoma. 2017 Joint Summits on Translational Bioinformatics, San Francisco, March 27 ~ 30, 2017. 2017.

35. Greenfield A, Madar A, Ostrer H, Bonneau R. DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. PLOS ONE. 2010;5(10):e13397.

36. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1. Cancer cell. 2010;17(1):98.

37. Singan VR, Simpson JC. Implementation of the Rank-Weighted Co-localization (RWC) algorithm in multiple image analysis platforms for

quantitative analysis of microscopy images. Source code for biology and medicine. 2016;11:2.

38. Melak T, Gakkhar S. Maximum flow approach to prioritize potential drug targets of Mycobacterium tuberculosis H37Rv from protein-protein interaction network. Clinical and translational medicine. 2015;4(1):61.

39. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic acids research. 2008;36(Database issue):D684-8.

40. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36(Database issue):D901-6.

41. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017;16(1):19-34.

42. McAdoo SP, Reynolds J, Bhangal G, Smith J, McDaid JP, Tanna A, et al. Spleen tyrosine kinase inhibition attenuates autoantibody production and reverses experimental autoimmune GN. Journal of the American Society of Nephrology : JASN. 2014;25(10):2291-302.

43. Lee JH, Kim DG, Bae TJ, Rho K, Kim JT, Lee JJ, et al. CDA: combinatorial drug discovery using transcriptional response modules. PloS one. 2012;7(8):e42573.

44. Zhong Y, Chen EY, Liu R, Chuang PY, Mallipattu SK, Tan CM, et al. Renoprotective effect of combined inhibition of angiotensin-converting enzyme and histone deacetylase. Journal of the American Society of Nephrology : JASN. 2013;24(5):801-11.

45. Sen B, Peng S, Tang X, Erickson HS, Galindo H, Mazumdar T, et al. Kinase-impaired BRAF mutations in lung cancer confer sensitivity to dasatinib. Sci Transl Med. 2012;4(136):136ra70.

46. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862-8.

47. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980-5.

48. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen — The Clinical Genome Resource. New England Journal of Medicine. 2015;372(23):2235-42.

49. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49(2):170-4.