**INFERENCE OF PERSONALIZED DRUG TARGETS VIA NETWORK PROPAGATION**

Ortal SHNaps[[1]](#footnote-1)

*School of Computer Science, Tel Aviv University  
Tel Aviv 69978, Israel  
Email: ortalits@gmail.com*

Eyal perry­

*School of Computer Science, Tel Aviv University  
Tel Aviv 69978, Israel  
Email: eyal.perry88@gmail.com*

DANA SILVERBUSH

*School of Computer Science, Tel Aviv University  
Tel Aviv 69978, Israel  
Email: dana.silverbush@gmail.com*

RODED SHARAN

*School of Computer Science, Tel Aviv University  
Tel Aviv 69978, Israel  
Email: roded@post.tau.ac.il*

This is where the abstract should be placed.

# Introduction

Precision medicine, an approach where medical treatment is tailored for a specific group of patients, is an arising paradigm in medical research and practice. Indeed, it is well known that some drugs affect only a specific sub-group of patients, while even harming other patients with the same disease (Poulikakos et al., 2010; Rothwell, 1995). In recent years, computational tools emerged to stratify diseases into informative subtypes (Hofree et al., 2013) and to predict sensitivity of subtype in order to optimally couple patients with existing medical treatments (Niepel et al., 2013).

However, the development of new treatments in the context of precision medicine is still scarce. Development of new drugs is an expensive and time consuming process; it takes about 15 years and up to 800 million dollars to convert a promising new compound into a marketed drug [2]. Consequently, there is increasing interest in computational prediction of drug targets (Chiang and Butte, 2009; Gottlieb et al., 2011; Hu and Agarwal, 2009; Lamb et al., 2006) used similarity among diseases to employ drugs designed for one disease to medicate another, as well as to prioritize new compounds as potential drugs. (Lamb et al., 2006) created a database containing ranked drug response gene expression profiles, allowing to query the database with a disease-specific genetic signature to identify drug response profiles that correlate to it. GBA (Chiang and Butte, 2009) predicts novel associations between drugs and diseases by assuming that if two diseases are treated by the same drug, alternative drugs treating only one of them might treat also the other. And finally, (Gottlieb et al., 2011) predicts novel associations between drugs and diseases by utilizing multiple drug–drug and disease–disease similarity measures for the prediction task. Some of the methods, such as (Gottlieb et al., 2011; Lamb et al., 2006) could be extended for personalized prediction of drugs, yet to this date efforts for personalized design of drugs had focused on experimental work (Zarrinkar et al., 2009) or small scale networks tailored for specific condition (Chuang et al., 2015; Ciaccio et al., 2015).

As drugs often act by inhibiting a target in a manner resembling a knockout, attempts were also made to predict candidates for drug targets by predicting the effect of potential knockouts. Aiming to explain and annotate protein-protein interactions (PPI) network, (Peleg et al., 2010) devised clustering methodologies to predict knockout effect in yeast. In-silico knockouts to infer drug target was examines by [Fatumo](http://www.ncbi.nlm.nih.gov/pubmed/?term=Fatumo%20S%5BAuthor%5D&cauthor=true&cauthor_uid=18313365) et.al. (Fatumo et al., 2009), by deleting reactions from a metabolic network to identify enzymes essential for the malaria parasite Plasmodium falciparum.

In this work we present a novel approach to tackle the drug target inference problem from a personalized perspective using in-silico knockouts based on propagation methods in a PPI network. Figure 1 provides an overview of the method: we start from a general PPI network and personal disease-related data. We rely on the framework described by Vanunu et al. (Vanunu et al., 2010) to prioritize casual genes by network propagation. We perform multiple network propagations in order to simulate the current patient state, the patient state after gene knockouts (by removing the gene's node from the network) and an estimated "healthy" state. We use these different states in order to rank the gene knockouts and retrieve a candidates list for potential novel drug targets.

The framework we presented is general and can be applied to any personalized disease-related data, with cancer being a pronounced candidate for application. Cancer is wildly heterogeneous, in that gene combinations can vary greatly between patients suffering from the same type of cancer. This is especially true in acute myeloid leukemia (AML), which has striking heterogeneity in gene mutations and expression aberrations across samples (Marcucci et al., 2011; 2013; Wang et al., 2013). We therefore evaluated our performance by applying it on patients suffering from AML, based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>. of mutated and differentially expressed genes (2013).

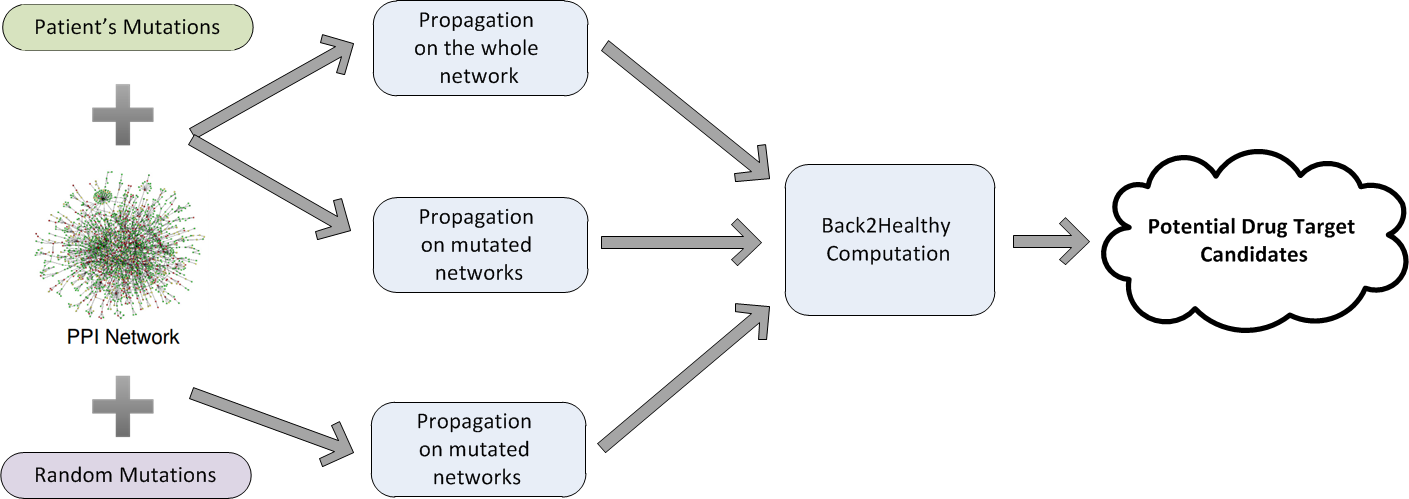


Fig. 1. An overview of the algorithmic pipeline.

The arrows are not clear: 1. "Propagation on the whole network" comes only from the network or from the +? "Propagation on mutated networks" comes from the +? Perhaps try to switch "PPI network" and "Patient's Mutations" and adjust the arrows and see if it becomes more clear. Also: B2H is ranking the random mutations? If so perhaps try to illustrate (for example you could add a "Ranked mutations" square in the same color as "Random Mutations".   
E&O: The arrows are all coming from the +s, that means that both PPI and mutation (either random or patient's) are required for each process (processes are blue; our first draft had explanation about the colors but it got lost through the revisions). About the B2H ranking, the B2H calculation uses the scores of the propagated random mutations and the scores of the propagated patient's mutations. In the figure we tried to show that the output of the 3 propagation processes is the input of the B2H calculation process.

# Results

We present a novel approach to tackle the drug target inference problem from a personalized perspective using in-silico knockouts in a PPI network. As described in Figure 1, we start from a general PPI network and personal disease-related data. We rely on the framework described by Vanunu et al. (Vanunu et al., 2010)to infer gene prioritization by network propagation. We perform multiple network propagations in order to simulate the current patient state, the patient state after gene knock-downs (by removing the gene's node from the network) and an estimated "healthy" state. We use these different states in order to rank the gene knockdowns and retrieve a candidates list for potential novel drug targets.

To evaluate our performance we applied our method to TCGA gene expression and mutation data of patients suffering from acute myeloid leukemia (AML). We gathered this dataset from the COSMIC cancer gene census (Forbes et al., 2011). First, we show that we can identify common AML causal genes by synthesizing the individual mutations set propagations and ranking according to propagation scores. Second, we show that by integrating results from a personalized knockout process we can infer potential drug targets.

We executed the algorithm using different settings for its alpha parameter (0.5, 0.75 and 0.9, see Methods) and the prior set of genes. To evaluate the results, we used three sets of known causal genes, varying in confidence and size: 10 AML causal genes from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (<http://www.genome.jp/kegg/>) (Kanehisa and Goto, 2000; Kanehisa et al., 2014), 94 AML causal genes from COSMIC (72 of which are in our PPI network) and 533 cancer causal genes from COSMIC (363 in the network). The application of the method to each patient resulted in a propagation score for each gene. We aggregated the rank of each gene over a random sample of 100 patients to yield a gene-based score, retaining the top 10% scoring genes. We then computed the hypergeometric enrichment of this set of genes with the different sets of known causal genes. All choices of resulted in significant p-values (), though the best one out of the three is 0.9, as shown in Figure 2A. This value was used in the sequel.

Dana: What P did you use for this test? E&O: we used the 4th option, we're not sre how to present it before the next paragraph.

Next, we wish to assess different prior knowledge (P) used to initiate the propagation (see Methods) gene sets based on the patients’ data. Each patient holds information about each gene - whether it is mutated and/or differentially expressed. Therefore, we examined four settings – defining P based on (i) mutated genes; (ii) differentially expressed genes; (iii) both, but running them separately and averaging the propagation scores obtained; and (iv) same as (iii) but taking the maximum scores rather than the average. We apply the method to each patient starting from the respected prior knowledge variant and aggregate the results to retain the top 10% affected genes according to the current variant. We again compute the hypergeometric enrichment of this set of genes with the different sets of known causal genes. All prior knowledge variants resulted in significant p-values (). The best variant was the first – setting P to be the set of mutated genes in each patient (Figure 2B), a choice we use in the sequel.

Fig. 2. Performance evaluation under different parameter (A) and prior knowledge set (B) choices. The red line stands for a p-value of .

The previous results imply that our propagation based scores are able to infer disease-related genes. We hypothesized that good drug targets for the disease could be genes whose knockout is predicted to reverse the disease-related effects. To identify such genes in-silico, we rerun the propagation based scoring while removing each gene in turn from the network, assessing the similarity between the obtained scores and those that characterize a “healthy” state. To this end, we use a Back2Healthy distance score (B2H; See Methods), and taking the top scoring genes as our best candidates for potential personalized drug targets.

The process above infers drug targets for each patient. As information about personalized drug targets is very scarce and hard to validate, we aggregated the results over all patients, evaluating the results using known AML drug targets derived from the DrugBank database (Knox et al., 2011; Law et al., 2013; Wishart et al., 2006, 2008) and COSMIC (see Methods). The top 10% scoring genes were highly enriched with known drug targets from both sources (Figure 3A, DrugBank: , COSMIC: ). To assess the personalized approach we took, we generated a "meta-patient", using consensus (appearing in at least 5 patients) mutated and differentially expressed genes derived by aggregating all AML patients. The results were insignificant (Figure 3B), underscoring the utility of a personalized approach.

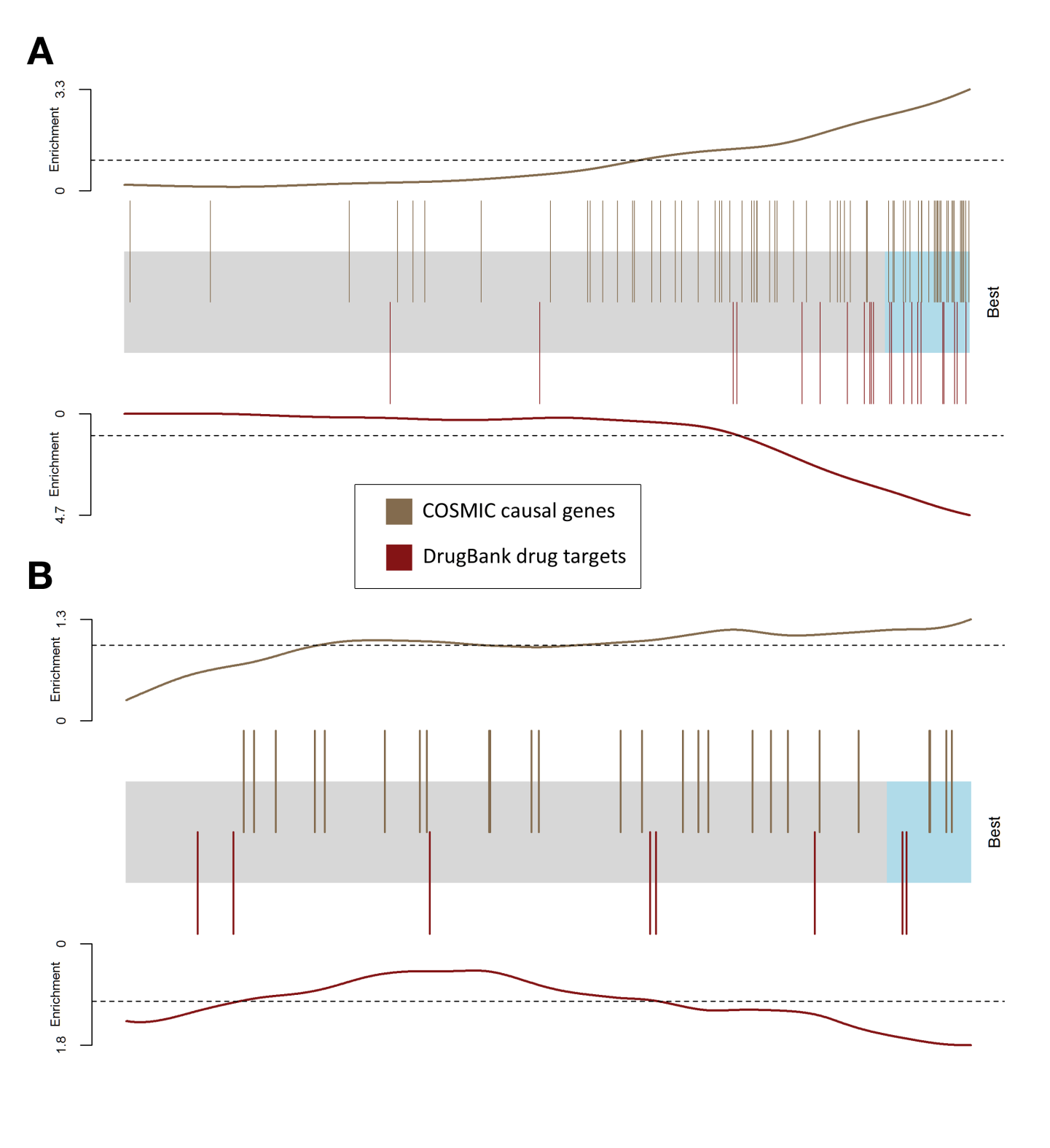


Fig. 3. Performance of drug target prediction. The candidate genes are represented by a shaded rectangle, where the top 10% are shaded cyan. Every overlaid bar stands for a single gene in a collection of known or potential drug targets. Traces above/below the bar represent relative enrichment. (A) The barcode plot was generated by running our method on each AML patient independently and aggregating the results. (B) The barcode plot was generated by running a similar single pipeline on the meta-patient.

# Methods

## Computing propagation score

We use the network propagation method described in Vanunu et al. (Vanunu et al., 2010). The input consists of a network with V as the set of proteins, E as the set of their interactions, represents the reliability of the interaction between and , and a prior knowledge protein set . Our goal is to prioritize the proteins in with respect to . We do so by defining a function that is both smooth over the network and accounts for the prior knowledge about each node.

Regarding our framework, the set of protein-protein interactions and their reliability are taken from the HIPPIE network [16]. The prior knowledge set is derived from the patient's mutation data.

As described by Vanunu et al. we use Laplacian normalization to produce the normalized network edge weight . Briefly we construct a sparse matrix from the edge weights , and construct a diagonal matrix with . The normalized edge weight matrix is computed as .

We define a prior knowledge function such that:

With the normalized weight matrix and the prior knowledge function , we use the iterative procedure described by Zhou et al. [20] to compute . Namely, starting with , we update at iteration as follows:

The procedure is repeated iteratively until convergence; namely we stop the iterations when the following condition occurs:

The propagation score for each gene is its rank among all genes after propagating the network, where lower ranks means higher value. In case of ties the ranks of the genes is averaged.

## Back2Healthy distance score

Let , be vectors of propagation scores for a chosen gene set (here, the set of differentially expressed genes of some patient) , where was generated by propagating on the original PPI network, while was generated by propagating on a “knockout” network, where one of the genes was removed. We define the Back2Healthy (B2H) distance between and as follows:

Let *k* be the size of the prior gene set of the patient (the patient’s set of mutated genes). For , we generate a score vector for by propagating the original PPI network and setting the prior knowledge set to be random nodes (disjoint from *A*) in order to simulate a “healthy” distribution of propagation scores for .

Next, for , define

Hence, represents the quantile of in our simulated distribution, and similarly for . Finally, is defined as:

# Conclusions

# References

Zhou, D.S. Wishart, DrugBank 4.0: shedding new light on drug metabolism, Nucleic Acids Res. Chiang, A.P., and Butte, A.J. (2009). SYSTEMATIC EVALUATION OF DRUG-DISEASE RELATIONSHIPS TO IDENTIFY LEADS FOR NOVEL DRUG USES. Clin. Pharmacol. Ther. *86*, 507–510.

Chuang, R., Hall, B.A., Benque, D., Cook, B., Ishtiaq, S., Piterman, N., Taylor, A., Vardi, M., Koschmieder, S., Gottgens, B., et al. (2015). Drug Target Optimization in Chronic Myeloid Leukemia Using Innovative Computational Platform. Sci. Rep. *5*.

Ciaccio, M.F., Chen, V.C., Jones, R.B., and Bagheri, N. (2015). The DIONESUS algorithm provides scalable and accurate reconstruction of dynamic phosphoproteomic networks to reveal new drug targets. Integr. Biol. Quant. Biosci. Nano Macro *7*, 776–791.

Fatumo, S., Plaimas, K., Mallm, J.-P., Schramm, G., Adebiyi, E., Oswald, M., Eils, R., and König, R. (2009). Estimating novel potential drug targets of Plasmodium falciparum by analysing the metabolic network of knock-out strains in silico. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. *9*, 351–358.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945–D950.

Gottlieb, A., Stein, G.Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. *7*, 496.

Hofree, M., Shen, J.P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. Nat. Methods *10*, 1108–1115.

Hu, G., and Agarwal, P. (2009). Human Disease-Drug Network Based on Genomic Expression Profiles. PLoS ONE *4*.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. *42*, D199–D205.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2011). DrugBank 3.0: a comprehensive resource for “omics” research on drugs. Nucleic Acids Res. *39*, D1035–D1041.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K.N., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929–1935.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., et al. (2013). DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. gkt1068.

Marcucci, G., Haferlach, T., and Döhner, H. (2011). Molecular Genetics of Adult Acute Myeloid Leukemia: Prognostic and Therapeutic Implications. J. Clin. Oncol. JCO.2010.30.2554.

Niepel, M., Hafner, M., Pace, E.A., Chung, M., Chai, D.H., Zhou, L., Schoeberl, B., and Sorger, P.K. (2013). Profiles of Basal and Stimulated Receptor Signaling Networks Predict Drug Response in Breast Cancer Lines. Sci. Signal. *6*.

Peleg, T., Yosef, N., Ruppin, E., and Sharan, R. (2010). Network-Free Inference of Knockout Effects in Yeast. PLoS Comput Biol *6*, e1000635.

Poulikakos, P.I., Zhang, C., Bollag, G., Shokat, K.M., and Rosen, N. (2010). RAF inhibitors transactivate RAF dimers and ERK signalling in cells with wild-type BRAF. Nature *464*, 427–430.

Rothwell, P.M. (1995). Can overall results of clinical trials be applied to all patients? The Lancet *345*, 1616–1619.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. *6*, e1000641.

Wang, H., Hu, H., Zhang, Q., Yang, Y., Li, Y., Hu, Y., Ruan, X., Yang, Y., Zhang, Z., Shu, C., et al. (2013). Dynamic transcriptomes of human myeloid leukemia cells. Genomics *102*, 250–256.

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. *34*, D668–D672.

Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. *36*, D901–D906.

Zarrinkar, P.P., Gunawardane, R.N., Cramer, M.D., Gardner, M.F., Brigham, D., Belli, B., Karaman, M.W., Pratz, K.W., Pallares, G., Chao, Q., et al. (2009). AC220 is a uniquely potent and selective inhibitor of FLT3 for the treatment of acute myeloid leukemia (AML). Blood *114*, 2984–2992.

(2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N. Engl. J. Med. *368*, 2059–2074.

1.  These authors contributed equally to this work. [↑](#footnote-ref-1)