Title: CELL FITNESS IS AN OMNIPHENOTYPE

Authors: Nicholas C. Jacobs[1,3], Ji Woong Park[1,2], Timothy R. Peterson[1,4]*

Affiliations: [1] Department of Internal Medicine, Division of Bone & Mineral Diseases, Department of Genetics, Institute for Public Health, [2, 3] Division of Basic and Biomedical Sciences: Computational and Systems Biology [2] and Molecular and Cellular Biology programs [3], Washington University School of Medicine, BJC Institute of Health, 425 S. Euclid Ave., St. Louis, MO 63110, USA. [4] Bio-I/O, 4320 Forest Park Ave., St. Louis, MO 63108, USA.

* **Correspondence to:** timrpeterson@wustl.edu (T.R.P).

ABSTRACT

Genotype-phenotype relationships are at the heart of biology and medicine. Numerous advances in genotyping and phenotyping have accelerated the pace of disease gene and drug discovery. Though now that there are so many genes and drugs to study, it makes prioritizing them difficult. Also, disease model assays are getting more complex and this is reflected in the growing complexity of research papers and the cost of drug development. Herein we propose a way out of this arms race. We argue for synthetic interaction testing in mammalian cells using cell fitness – which reflect changes in cell number that could be due to a number of factors – as a readout to judge the potential of a genetic or environmental variable of interest (e.g., a gene or drug). That is, if a mammalian gene or drug of interest is combined with a known perturbation and causes a strong cell fitness phenotype relative to that caused by the known perturbation alone, this justifies proceeding with the gene/drug in more complex models like mouse models where the known perturbation is already validated. This recommendation is backed by the following: 1) human protein-coding genes important to cell fitness under normal growth conditions involve nearly all classifications of cellular and molecular processes; 2) Nearly all human genes important in cancer – a disease defined by altered cell number – are also important in other complex diseases; 3) Many drugs affect a patient's condition and the fitness of their cells comparably. Taken together, these findings suggest cell fitness could be a broadly applicable phenotype for understanding gene and drug function. Measuring cell fitness is robust and requires little time and money. These are features that have long been capitalized on by pioneers using model organisms that we hope more mammalian biologists will recognize.

INTRODUCTION

Biological model systems have made numerous contributions to human health, which is notable because often they were used without regard to their eventual applications. For example, many transformative medicines and field-changing paradigms exist because of discoveries in organisms such as yeast and worms (1-4). Still, some argue that even mice aren't good models of disease. Some researchers are regardless pushing forward and focusing on creating new ways to study human disease using model organisms (5, 6). Yet, many important human genes aren't conserved in model organisms. Moreover, even if one uses a mammalian system, it is assumed that the context being studied, e.g., tissue type, matters. There is therefore a need for models that strike a balance between the ease and robustness of model organisms while preserving the signaling networks important to human disease.

Modeling disease has gained new relevance as the number of human population genetic studies have exploded (7). Thousands of genes are being implicated in human diseases across categories from cancer to autism, Alzheimer's, and diabetes. Among the surprises from these data are findings such as "cancer" genes – genes understood to be important in cancer – being linked to diseases seemingly unrelated to cancer. For example, one of the earliest genes found to be mutated in the behavioral condition, autism (8), is the well-known tumor suppressor gene, PTEN. On first pass, it might be difficult to glean what autism might have to do with the uncontrolled cell fitness that PTEN inactivation causes. However, it is clearer when one considers that patients with autism often have larger brains and more cortical cells (9). Nevertheless, the extent to which the various functions of genes like PTEN are relevant to disparate diseases begs the question of how we molecularly classify and model disease.

As with studying disease genes, new treatments have often suffered from biases that prevented us from seeing their value outside of the context in which they are originally characterized. For example, going back in history, one of the largest miscalculations by the pharmaceutical industry was the cholesterol-lowering statins being held back from the clinic because they were shown to be toxic to cultured mammalian cells. Those early findings scared off drugs companies because they suggested statins might not be safe for humans (10, 11). Merck understood these results better than their competitors and the statins went on to be one of the biggest success stories in drug development history. The successes of drug repositioning also tell us that drugs can function in multiple “unrelated” contexts. For example, malarial drugs, such as hydroxychloroquine (Plaquenil), are now used to treat rheumatoid arthritis and autoimmune disorders (12). There are also chemotherapeutics being considered for depression (13). Drug repositioning makes sense from a molecular perspective because many drug side effects could actually be on-target effects (14). To take it further, now with so much knowledge in the “-omic” era, it might be the expectation rather than the exception that a mechanistic target for a drug has many functions in the body. For example, the serotonin transporter, SLC6A4 (a.k.a. SERT), is the target of the widely prescribed antidepressants, Prozac and Zoloft. SLC6A4 is expressed at high levels in the intestines and lungs (15), and regulates gut motility and pulmonary blood flow (16, 17) in addition to mood. This demonstrates the need to find a better way to understand the diverse roles of drugs and their targets.

Herein, we address these aforementioned issues. We propose that synthetic interaction testing in mammalian cells using cell number/fitness as a readout is a relevant and scalable approach to understand surprisingly diverse types of molecular functions, diseases, and treatments. This proposal requires a change in thinking because synthetic interaction testing using human cells has historically almost always been discussed in the context of cancer (18, 19). This makes sense considering cancer has become defined by alterations in several parameters that

contribute to cell number/fitness, e.g., proliferation and apoptosis (20, 21). However, we argue fitness-based synthetic interaction testing has a much more general purpose. We provide evidence that cell fitness is a generalizable phenotype because it is an aggregation of phenotypes. To the extent that it might be an aggregation of all possible phenotypes – an omniphenotype – suggests its potential as a pan-disease model for biological discovery and drug development.

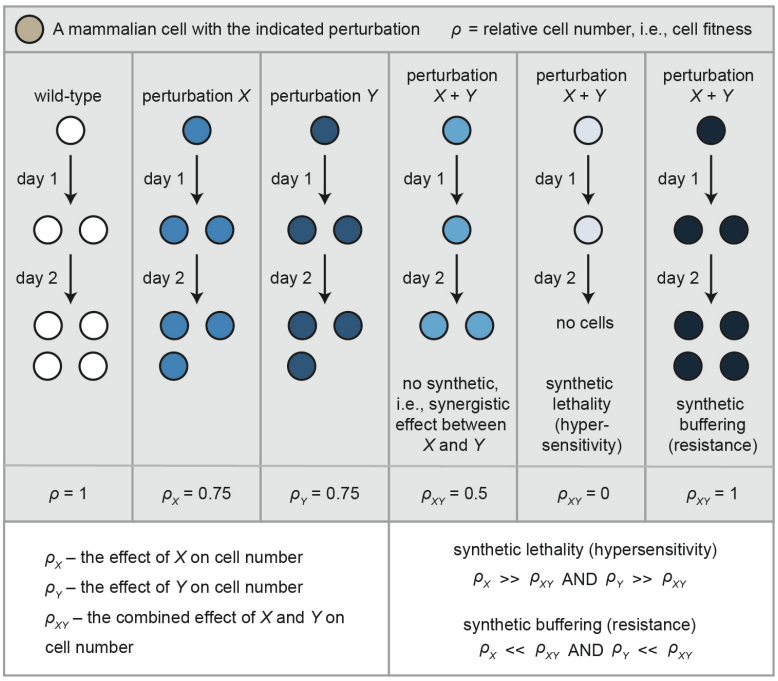

Figure 1: A current, commonly used framework for synthetic interaction testing. A synthetic interaction between two perturbations, *X* and *Y*, means the effect of *X* combined with *Y* on cell fitness is significantly greater than the effect of either alone. Synthetic interactions can cause lethality, a.k.a., hypersensitivity, or can be buffering, a.k.a., confer resistance. Both are depicted mathematically in the figure where ρ – the phenotype – is relative cell number.

RESULTS

Nearly all cell processes affect cell fitness. A simple way to understand synthetic interactions is to consider two genes. It is said there is a synthetic interaction between two genes if the combined effect of mutating both on cell fitness, heretofore denoted by "ρ" (as in phenotype), is much greater than the effects of the individual mutations (Fig. 1). If there is an interaction, it can

be either be a negative synergy, where the double mutant cells grow much worse (a.k.a., synthetic lethality), or a positive synergy, where the cells grow much better (a.k.a., synthetic buffering), respectively, than the individual mutant cells (Fig. 1).

Amongst the ~400 million possible synthetic interactions between protein coding human genes, less than 0.1% have been mapped to date (22). It is challenging to scale synthetic interaction testing in human cells both because of the size of the experiments needed and because sometimes neither interactor has a known function in the ultimate biological context of interest (23). Therefore, we sought to develop a framework that would allow us to infer synthetic interactions at scale as well as leverage interactors of known importance to explain interactors of unknown significance. The convention of characterizing an unknown in the context of a known is a typical approach taken by biologists for making discoveries. To formalize our approach, we heretofore refer to the known interactor in a synthetic interaction pair as *K,* and the unknown as *U.* With this convention, *U* is what's new and interests the researcher, whereas *K* is a factor that's known in the field. *U* and *K* could represent a gene, disease, or drug.

Several recent works identified "essential" genes – genes that are required for cell fitness (24-26) – and we wondered what gene ontology (GO) pathways they participate in. We applied the principal of the minimal cut set (MCS), which is an engineering term used by biologists to mean the minimal number of deficiencies that would prevent a signaling network from functioning (27, 28). MCS fits well with the concept of synthetic lethality (18), and this was recognized by Francisco Planes and colleagues to identify RRM1 as an essential gene in certain multiple myeloma cell lines (27).

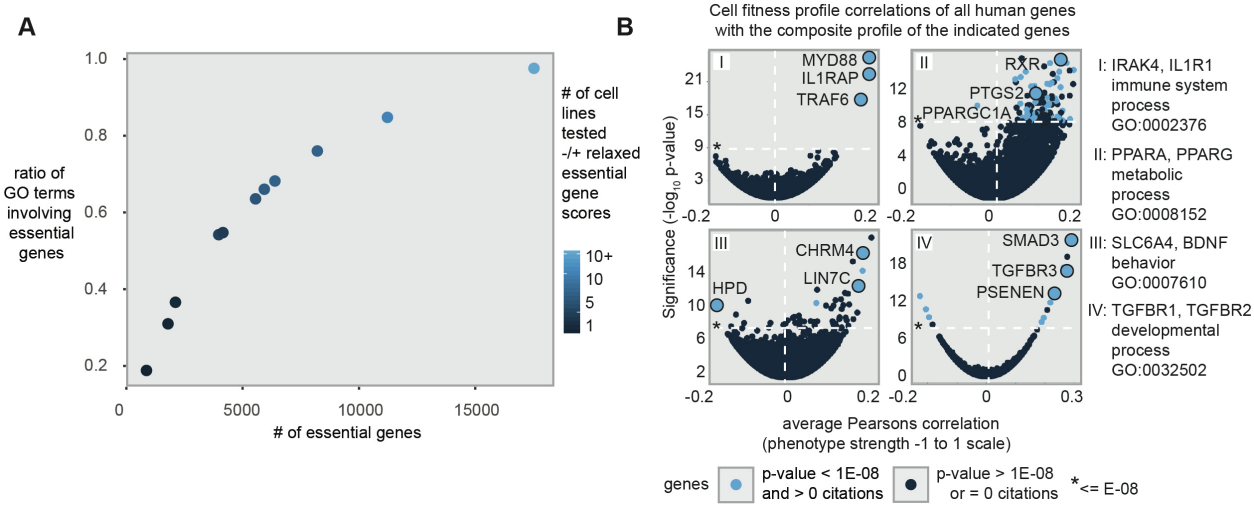

Figure 2. Nearly all GO processes are essential for cell fitness. (A) The y-axis is the ratio is the number of distinct GO terms linked to the essential genes in the tested cell lines divided by the total number of distinct GO terms for all tested human genes (21,246). The x-axis is the number of essential genes determined for the indicated number of cell lines. + indicates relaxed, i.e., $p < 0.10$, cell fitness scoring. **(B)** Fitness profile correlations across 563 human cell lines of mutant cells of genes from the same GO categories. The statistically significant genes that have been most highly co-cited with the indicated genes are color-coded light blue.

We generalized the idea of MCS to hypothesize that different cell types might be relatively defective in components of specific GO pathways such that if a gene in one of these pathways was knocked out, it would produce a synthetic lethal interaction with those components (for clarity throughout this study, by “genes”, we refer to human protein coding genes). In this case, the synthetic interaction testing is as follows: the known, *K,* is the gene that is essential in at least one cell type and *U* is the unknown, synthetically-interacting GO pathway of interest. Indeed, we found that essential genes are involved in diverse GO processes and the number of GO processes they participate in scales with the number of cell lines examined (Fig. 2A, Supp. Table 1). For example, if one cell line is assessed, 2054 out of 21246 total genes assessed were essential and these essential genes involve 36% (6444 out of 17793) of all possible GO categories. In examining five cell lines, 3916 essential genes were identified that involve 54% of all GO categories (9577/17793). With ten cell lines, 6332 essential genes were identified that involve 68% of all GO categories (12067/17793). Therefore, essential genes involve the majority of GO processes. Because the Fig. 2A graph appears asymptotic, this suggests if enough cell

types are tested, potentially all GO processes would be essential. To extrapolate beyond the ten mammalian cell lines where essential genes have been systematically characterized, we relaxed the significance on the cell fitness scores, e.g., from $p < 0.05$ to $p < 0.10$. This relaxed list of fitness-affecting genes encompasses nearly all GO processes (Fig. 2A, 97.2%, 17291/17793 GO terms; 17473/21246 genes). Taken together, this suggests the majority if not all GO processes affect cell fitness.

Recently, genome-scale, protein coding gene inactivation screening using CRISPR has been performed on hundreds of human cell lines (29). Using this data and the principle of gMCS, we reasoned that the single gene mutant cell fitness profiles for genes in the same GO processes would strongly correlate across cell lines. We selected gene pairs from several diverse GO categories (immune system process (GO:0002376), behavior (GO:0007610), metabolic process (GO:0008152), and developmental process (GO:0032502)) and asked how the fitness profiles of other members of the same GO category would rank compared with all 17,634 fitness profiles that were assessed. Impressively, other genes from the same GO category and/or highly co-cited genes were the most correlated in their fitness profiles for each GO category we assessed (Fig. 2B, Supp. Table 1). This suggests cell fitness can predict gene function for disparate biological processes.

Nearly all common disease genes are cancer genes. The omnigenic model states that all genes are important to complex disease (30). Yet, there is a scale and some genes are more important than others. These relatively more important genes are referred to as “core” (a.k.a. “hub”) genes. We reasoned that the published literature with its millions of peer reviewed experiments would be a comprehensive and unbiased way to inform us on which genes are core genes for a given complex disease. At the same time, we were interested in leveraging the literature to identify genes that would most strongly synthetically interact with these core genes.

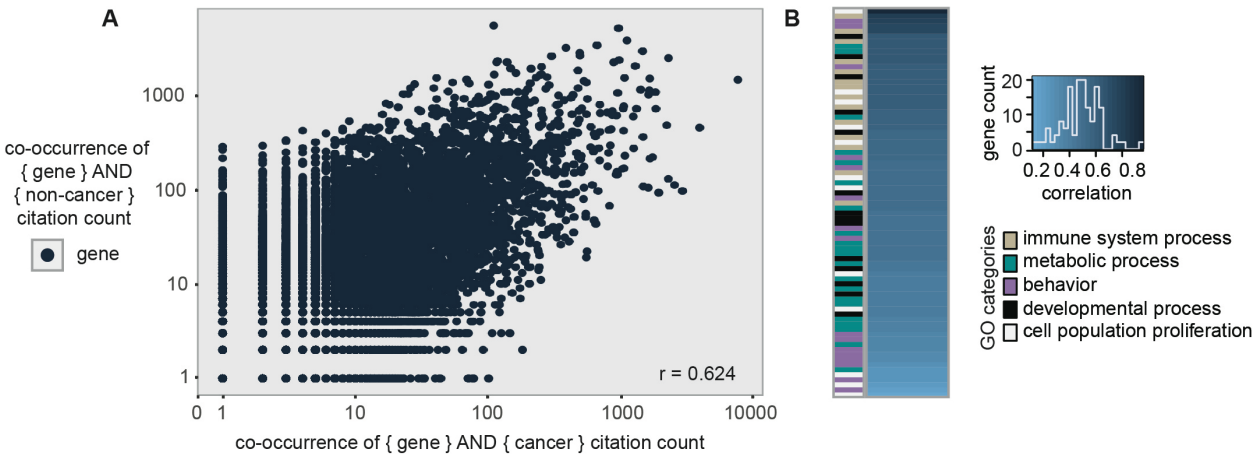

Figure 3: Nearly all common disease genes are cancer genes. (A) The co-occurrences of each human gene with cancer (x-axis) and non-cancer (y-axis) conditions in all currently existing ~30 million PubMed citations were counted. The non-cancer conditions include all citations not classified with the "neoplasm" MeSH classifier. It includes common conditions like infection, Alzheimer's, cardiovascular disease, diabetes, obesity, depression, inflammation, osteoporosis, hypertension, and stroke amongst many other MeSH classifiers. Line of best fit - Slope, 0.67; Correlation coefficient, r = 0.624. **(B)** Gene ontology (GO) classifications that comprise the non-cancer/cancer correlation in (A). Dark blue color represents a stronger correlating GO category, lighter blue represents a weaker correlating category.

Here we defined a synthetic interaction as follows: the gene(s) mentioned in a citation is the unknown, *U*. This is by convention because the gene(s) most often is not known in whatever context was studied, e.g., diabetes, and that's why it warrants a publication. Whereas all other published genes in that context are the known, *K*. To measure synthetic interaction strength between *U* and *K*, we assessed their co-occurrence with the term, "cancer" as a proxy for cell fitness because cancer is a complex disease defined by altered cell number. Specifically, we assessed all citations linked to the MeSH term "neoplasm" vs. those that were not (i.e., "cancer" vs. "non-cancer" citations as we will heretofore refer to them, see Methods for more details). For context, the non-cancer literature includes conditions like the top ten most common causes of death according to the World Health Organization (31) and other sources: Alzheimer's, cardiovascular disease, diabetes, obesity, depression, infection, osteoporosis, hypertension, stroke, and inflammation. In analyzing all currently available PubMed citations (~30M), we identified a compelling relationship: most genes are cited in the cancer literature to a similar extent as with the non-cancer literature (Fig. 3A, Supp. Table 1). This high correlation (r =

0.624) suggests for many (non-cancer) disease states, genes important to them can be identified using synthetic interaction testing. That some genes are cited much more than others across disease categories is consistent with the multiple functions of core/hub genes.

We analyzed the pattern we detected in Fig. 3A in more detail. To understand what GO categories are leading to better cancer/non-cancer citation correlations, we binned the cancer and non-cancer citations for all genes by their GO categories. We found no clear pattern of cancer-related GO categories or sub-categories that contribute most vs. least to the cancer/non-cancer citation correlation (Fig. 3B, Supp. Table 1). For example, the cancer-related GO category, cell population proliferation (GO:0008283), is interspersed amongst the same four diverse and less explicitly cancer-related GO categories we assessed in Fig. 2: immune system process, behavior, metabolic process, and developmental process. This lack of a pattern suggests that many more processes than might be expected contribute to cancer. That the GO sub-category, maintenance of cell number (GO:0098727), was amongst the top 10 most correlated GO sub-categories supports our use of cancer as a proxy for cell fitness (Supp. Table 1). If it is found that many, most, or all GO categories (and thus all genes) contribute to cancer, in future studies it will be interesting to classify each into one or more of the categories outlined in the well-known “Hallmarks of cancer” review by Weinberg and Hanahan (20, 21).

A

drug	human condition	citation (PMID)	IC50 range	citation(s) (PMID)
fluoxetine (SSRI)	Major Depressive Disorder	28291260	1-8μM	29094413
valproic acid lithium (GSK3 inh.)	Muscular Dystrophy	30006409 24091664	1-2mM 10-20mM	22469995 11337498
ruxolitinib (JAK inhibitor)	Down Syndrome	27472900	15-370nM	23560534 22875628
calmidazolium (calmodulin antag.	Alzheimer's Disease	12901840	6-7.1μM	3940184
ethanol	Alcoholism	25129674	7.5-17.6mM	20074269 15208159
dexamethasone (glucocorticoid)	Congenital Adrenal Hyperplasia	12519866	1-100μM	19421908 12230951
7DG (PKR inhibitor)	Cornelia De Lange Syndrome	26725122	300nM-7.5μM	28758979
simvastatin (statin)	Optic Neuritis	19948167	6-45μM	23675166
metformin (biguanide)	Diabetes Cancer	28173075	440μM-46.3mM	22576795
rapamycin (mTOR inhibitor)	immuno-suppressant	21067467	15pM to >100nM	8778023

patient drug response correlates with drug response on patient cells

drug IC50 value range on human cells or cell lines

B

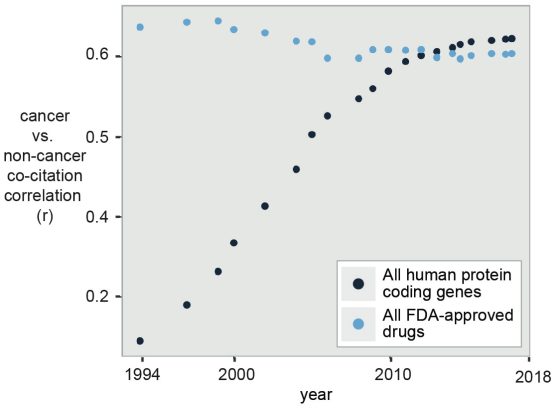

Figure 4. Establishing the efficacy of FDA-approved drugs using cell fitness. (A) Drugs where their effects on a patient condition and the fitness of the same patient's cells correlate. Listed drugs met two criteria: 1) Their effect on a patient's condition and the same patient's cell fitness were significantly correlated; 2) The drug had differing IC50 values – the concentration at which cell numbers are 50% of their untreated values – in multiple, distinct human cell types or cell lines. **(B)** Correlation (r) over time of human protein-coding genes or FDA-approved drugs co-cited with cancer or non-cancer conditions analyzed as in Fig. 3A.

***Drug effects on patient cell fitness are reflective of the patient condition response to drug*.** Like with mutations in genes, drugs can have different effects in different genetic backgrounds. We analyzed the published literature to identify reports where the fitness of patient-derived cells in response to a drug was also predictive of the same drug's effect on the patient's condition. In this case, *U* is the genetic background of the patient cells and *K* is the drug. What's notable in considering all the studies we identified is that cell fitness is predictive of patient drug response for widely divergent medications (Fig. 4A). These results might be surprising outside the context of cancer therapy, but they should not be because the level of growth inhibition (inhibitory concentration - IC50) for these drugs varies widely in different cell lines, which have different genetic/epigenetic backgrounds (Fig. 4A, Supp. Table 1). Note that this latter result is similar to that presented in Fig. 2A, though using chemical genetics rather than genetics.

To understand how FDA-approved drugs are used for cancer vs. non-cancer conditions and how their usage might change over time we analyzed PubMed as in Figure 3A. Like with PTEN

as previously discussed, our understanding of the role of many genes in cancer has preceded that of our understanding of their roles in other diseases. We reasoned that with several advances, such as various new –omic analyses, our understanding of the molecular basis of many diseases is "catching up" to cancer more recently. Thus, we expect the correlation between human protein coding genes and FDA-approved drugs being cited both in and out of the context of cancer to be increasing over time. Indeed, human genes are increasingly being cited with similar frequency in cancer and non-cancer disease contexts (Fig. 4B, Supp. Table 1). This supports the idea that one should be able to increasingly make use of the literature to understand an unknown gene's function. Like with human genes, we found that clinical drugs are commonly cited in both the cancer and non-cancer literature (correlation = 0.599), albeit this correlation hasn't increased over time like with genes (Fig. 4B, Supp. Table 1). This latter result is interesting in light of the increased focus on repositioning many non-cancer drugs for cancer (32) (e.g., immunosuppressant, rapamycin and anti-diabetic, metformin), and vice versa (e.g., chemotherapeutics for autism, senolytics for aging (33).

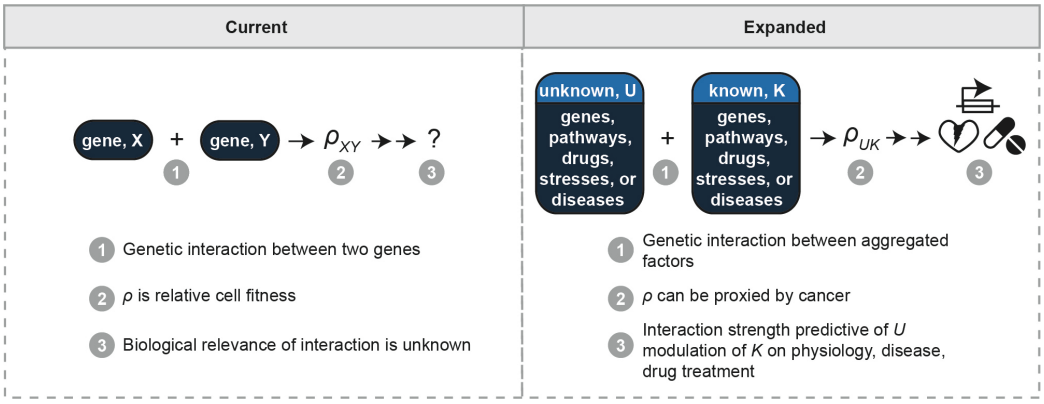

Figure 5. Current vs. expanded conceptual frameworks for synthetic interaction testing. A currently, commonly used framework for synthetic interaction testing is as follows: 1) Genetic interactions are tested between two genes, e.g., *X* and *Y*; 2) *f,* the phenotype being measured, is relative cell fitness; 3) It isn't known how the interaction strength between *X* and *Y* might be relevant to physiology or disease. The expanded framework described herein is as follows: 1) Genetic interactions can be tested between aggregated factors such as multiple genes and/or pathways (related to Figure 2); 2) *f* can be proxied by the relative involvement of the interactors in cancer (related to Figure 3); 3) One interactor, *K*, has a known effect on a physiology or disease of interest. The effect of the other, *U*, in that context can therefore be measured relative to *K* (related to Figure 4). The implications of mammalian cell fitness-based synthetic interaction testing as described here is that the results using cell fitness can be readily extrapolated to the ultimate phenotype of interest, e.g., osteoclast resorptive ability. That is, the magnitude of the synergy on cell fitness of the two factors under consideration is expected to be proportional to the magnitude of their synergy on the phenotype of interest.

DISCUSSION

We propose cell fitness as a simple, yet comprehensive approach to rank less-studied genes, cell processes, diseases, and therapies in importance relative to better-studied ones (Fig. 5). Rather than needing ever more refined experiments, this work suggests the antithetical approach. That is, one can use fitness in mammalian cells to estimate the importance of a poorly characterized factor they are interested in (herein referred to as *U*) in a biological context of interest by comparing the new factor's individual effect on cell fitness vs. its combined effect with a factor with a known role in the biological context (herein referred to as *K*) (Fig. 5). We recently published a proof-of-concept of this idea – which can be thought of as an expanded definition of synthetic interaction testing – with the osteoporosis drugs, nitrogen-containing bisphosphonates (N-BPs) (22, 34, 35). In this work, we identified a gene network, *ATRAID-SLC37A3-FDPS*, using cell-fitness-based, drug interaction screening with the N-BPs, and subsequently demonstrated this network controls N-BP responses in cells, in mice, and in humans (22, 34, 35). Besides this and the statins, there are numerous other examples in the literature such as with the mTOR pathway (36) where the steps from cell fitness to phenotypes relevant in people can be readily traced.

Cell fitness is a phenotype that has long had traction in the yeast and bacteria communities, but not yet in mammalian systems outside of cancer. To increase awareness of its utility, below we address its relevance as a phenotype to the different levels at which mammalian biologists tackle their work: at the level of cell and molecular processes, at the level of disease, and at the level of environmental factors such as drugs.

Relevance to cell and molecular processes (Fig. 2). Our Figure 2 results suggest one can study the majority of cell and molecular processes using cell fitness-based, synthetic interaction

testing. This is useful for molecular biologists because we often spend considerable time developing experimental models. For example, large efforts are often spent upfront (before getting to the question of interest) developing technology, such as cell types to model various tissues. This is done because of the often unquestioned assumption that the cell type being studied matters. While this can be true, it is worth considering whether it's the rule or the exception. In support of the latter, in single cell RNAseq analysis, 60% of all genes are expressed in single cells and the percentage jumps to 90% when considering 50 cells of the same type (37). Similar percentages are obtained at the population level when comparing cell types (38, 39). This argues that the majority of the time, there should be a generic cell context to do at least initial "litmus" synthetic interaction testing of a hypothesis concerning a new genetic or environmental factor of interest before proceeding with a more complex set-up. Like with cellular phenotypes, many molecular phenotypes such as reporter assays require domain expertise, can be hard to elicit, and might not be easily reproducible. Whereas cell fitness is robust and readily approximated by inexpensive assays such as those to measure ATP levels.

Relevance to diseases (Fig. 3). Our PubMed analysis lends growing support to the use of the published literature as a data source (40). The costs to perform the wet-lab experiments required to publish in the "-omic" era aren't small and favor well-resourced teams. Gleaning insights from the existing literature, which is large and growing rapidly, but at the same time knowable especially when approached computationally, is an underappreciated way to level the playing field.

One counterintuitive insight from Figure 3 is that core genes on the diagonal might not be as good drug targets as genes off the diagonal because the core genes are involved in more diseases. This might mean that because of their importance, core genes might be too networked with other genes to be targeted by therapies without the therapies causing off-target

effects. Fortunately, this might then also mean those genes off the diagonal might make more appealing drug targets because they are less networked. Thus, in this case we would use cell fitness as a filter to prioritize genes for further study. This is important because like with income inequality where the “rich get richer” certain genes get cited over and over (23), and approaches to aggregate data are known to favor the “discovery” of already well-studied genes over less studied genes (41). When applied in unbiased methods such as with genome-wide CRISPR-based screening (22), we expect our approach to be particularly impactful in addressing this issue of identifying important genes that historically are ignored.

Relevance to environmental factors, e.g., drugs (Fig. 4). With the exception of oncology, it could be argued pharmacogenomics is only slowly going mainstream in many fields. One reason for this is the poor feasibility and high cost-to-benefit ratio of current genetic testing strategies to predict drug response (42). Though more work needs to be done to test the generalizability and “real-world” application of the studies we highlight in Figure 4A, they do provide important initial evidence for using cell fitness as a diagnostic assay for precision medicine. In addition to drug diagnostics, these studies have implications for drug discovery. Both target-based and phenotypic-based drug discovery have drawbacks (43). Target-based approaches don’t necessarily tell you about the phenotype and phenotypic screens don’t tell you about the target. On the other hand, cell fitness-based, drug response genetic screening (44) can potentially be the best of both worlds because it can extrapolate well to the phenotype of interest and provides the gene target in the same screen.

Still, we acknowledge limitations in our findings. Our work focuses on genes and drugs as the genetic and environmental variables of interest. Future work must determine to what extent our work would apply to other variables, such as epigenetic signatures. Also, many genes aren’t published on due to human factors, which bias our PubMed-based analyses. As more methods

get developed and more knowledge accumulates, a scientist's job of figuring out where to focus their attention gets harder. Our view is that cell fitness could be used more often to help scientists triage their opportunities.

AUTHOR CONTRIBUTIONS

T.R.P, N.C.J, and J.P conceptualized the project. N.C.J and T.R.P performed the analysis. T.R.P wrote the paper and N.C.J and J.P edited it.

FUNDING

This work was funded by NIH grants, K99/R00AG047255 and R01AR073017, AWS Research Credits, and Washington University School of Medicine startup funding to T.R.P. Though there is no pending patent applications or financial transactions based on this work, T.R.P acknowledges potential future financial conflicts of interest as a major shareholder in the St. Louis-based, biotechnology company, Bio-I/O, which conducts research related to this work.

METHODS

Code. All code created for this work is available at: https://github.com/tim-peterson/omniphenotype.

Gene ontology and essential gene analysis (Fig. 2A). A list of all human genes was obtained from NCBI, ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/. Gene ontology (GO) information for each human gene was obtained from BioMart, https://useast.ensembl.org/info/data/index.html. Essential gene data was obtained from Blomen et al., Wang et al., and Hart et al. (24-26). There were ten cell lines tested: K562, KBM7, Jiyoye CS, Raji CS, HAP1, HCT116, DLD1, HeLa, GBM, RPE1. The code used to intersect the GO terms for each essential gene is available at the aforementioned Omniphenotype Github repository. Gene ontology: Developmental process (6297 genes),

For the relaxed filter data point, we used fitness scores at $p < 0.1$ for both the Wang and Hart studies. For the Blomen study, p-values > 0.05 weren't given, therefore we took all genes with fitness scores less than 0.5 standard deviations above the mean.

Gene inactivation fitness profile correlation (Fig. 2B). Single gene inactivation cell fitness profiles from the Broad Institute and Sanger Institute were downloaded from the DepMap website, https://depmap.org/portal/download/ using the 2019q2 dataset. Fitness profiles for singe genes from the same GO category were compared pairwise using the python pearsonr() function as detailed in the aforementioned Omniphenotype Github repository.

PubMed analysis (Fig. 3 and 4). In both Fig. 3 and 4, PubMed was analyzed using the E-utilities (https://www.ncbi.nlm.nih.gov/books/NBK25497/) API endpoint: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi. The results from these queries were either PubMed IDs (PMIDs), which were counted (Fig. 3A-B, 4B), or abstracts, which were manually analyzed for relevant information (Fig. 4A). The code for both analyses is available at aforementioned Omniphenotype Github repository.

Cancer vs. non-cancer analysis (Fig. 3A). All PMIDs for each human gene and their homologs and all PMIDs for each Medical Subject Headings (MeSH, https://www.nlm.nih.gov/mesh/meshhome.html) were collected into tab-delineated files. A citation was determined to be associated with a particular MeSH term if it came up from searching PubMed with "<MeSH term> [mesh]". A citation was determined to be associated with a particular gene if it has been labeled in "Related articles in PubMed" within the Gene resource of NCBI. See Omniphenotype Github repository for how these application programming interface (API) calls were made. The gene-PMID file was intersected with the MeSH-PMID file

and the resulting gene-MeSH term list was separated by the MeSH term “neoplasm”. Meaning, all citations that were annotated as referring to a gene that were co-annotated with the MeSH term “neoplasm” were considered a “cancer” citation and all other citations that were co-annotated as referring to a gene and any other MeSH term besides “neoplasm” were considered “non-cancer” citations. Those citations which mapped to both neoplasm as well as other conditions were considered “cancer” citations. Genes with less than 10 citations were excluded from the analysis. The correlation co-efficient of cancer vs. non-cancer citation counts for all genes was calculated using R using the lm() function on log-log data.

Gene Ontology-subsetted, cancer vs. non-cancer citation analysis (Fig. 3B). To determine the diversity of correlations for cancer vs. other conditions among GO categories, we took a subset of major GO categories (immune system process, metabolic process, behavior, developmental process, and cell population proliferation) as a representation of a diverse subsampling of biological processes. We then iterated through the sub-category/child terms of each of these major GO categories and found the correlation of all cancer vs. non-cancer citations for genes associated with that term. The results are displayed as a 1-Dimensional heatmap with a secondary mapping of the higher order GO categories also displayed.

Drug response in patients vs. patient cells analysis (Fig. 4A). To identify studies that reported findings on patient cell drug responses that correlated with the patient, the phrase “human lymphoblastoid cell lines proliferation” was queried in PubMed and 804 abstracts were returned. “PMBC” as in peripheral mononuclear blood cells was substituted for lymphoblastoid cell lines (LCLs) to identify the simvastatin study. The LCL and PMBC citations were manually curated to identify relevant citations that mentioned drug responses on cell proliferation. The IC50 studies were obtained by querying [drug] AND “cell viability” OR “IC50”.

Co-citation time series analysis (Fig. 4B). PubMed PMIDs are largely chronological, i.e., higher PMIDs mostly mean a more recent citation and vice versa, with the exception of PMIDs < 8M as well as those between 12M and 15M. As of July 9, 2019, citations up to the end of 2017 have been annotated with MeSH terms and gene associations such that cancer vs. non-cancer correlations for multiple time points could be determined. Citations were split every million PMIDs and for the purposes of graphing the results the calendar year was approximated. The FDA-approved drug list was obtained from: https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files. Similar to the cancer vs. non-cancer analysis for genes, all PMIDs for each drug were collected into a tab-delineated file, intersected with the MeSH-PMID file, and the resulting list was separated by those citations that were co-annotated with the MeSH term “neoplasm” and those that were not.

REFERENCES

1. Heitman J, Movva NR, Hall MN. Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast. Science. 1991;253(5022):905-9. PubMed PMID: 1715094.
2. Takeshige K, Baba M, Tsuboi S, Noda T, Ohsumi Y. Autophagy in yeast demonstrated with proteinase-deficient mutants and conditions for its induction. J Cell Biol. 1992;119(2):301-11. PubMed PMID: 1400575; PMCID: PMC2289660.
3. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998;391(6669):806-11. doi: 10.1038/35888. PubMed PMID: 9486653.
4. Spector JM, Harrison RS, Fishman MC. Fundamental science behind today's important medicines. Sci Transl Med. 2018;10(438). doi: 10.1126/scitranslmed.aaq1787. PubMed PMID: 29695453.
5. Rodriguez TP, Mast JD, Hartl T, Lee T, Sand P, Perlstein EO. Defects in the Neuroendocrine Axis Contribute to Global Development Delay in a Drosophila Model of NGLY1 Deficiency. G3 (Bethesda). 2018. doi: 10.1534/g3.118.300578. PubMed PMID: 29735526.
6. Sirr A, Scott AC, Cromie GA, Ludlow CL, Ahyong V, Morgan TS, Gilbert T, Dudley AM. Natural Variation in SER1 and ENA6 Underlie Condition-Specific Growth Defects in Saccharomyces cerevisiae. G3 (Bethesda). 2018;8(1):239-51. Epub 2017/11/16. doi: 10.1534/g3.117.300392. PubMed PMID: 29138237; PMCID: PMC5765352.
7. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2018;19(2):110-24. doi: 10.1038/nrg.2017.101. PubMed PMID: 29225335.
8. Knafo S, Esteban JA. PTEN: Local and Global Modulation of Neuronal Function in Health and Disease. Trends Neurosci. 2017;40(2):83-91. doi: 10.1016/j.tins.2016.11.008. PubMed PMID: 28081942.
9. Courchesne E, Mouton PR, Calhoun ME, Semendeferi K, Ahrens-Barbeau C, Hallet MJ, Barnes CC, Pierce K. Neuron number and size in prefrontal cortex of children with autism. JAMA. 2011;306(18):2001-10. doi: 10.1001/jama.2011.1638. PubMed PMID: 22068992.
10. Endo A. The discovery and development of HMG-CoA reductase inhibitors. 1992. Atheroscler Suppl. 2004;5(3):67-80. doi: 10.1016/j.atherosclerosissup.2004.08.026. PubMed PMID: 15531278.
11. Endo A. A historical perspective on the discovery of statins. Proc Jpn Acad Ser B Phys Biol Sci. 2010;86(5):484-93. Epub 2010/05/15. doi: 10.2183/pjab.86.484. PubMed PMID: 20467214; PMCID: PMC3108295.
12. Plantone D, Koudriavtseva T. Current and Future Use of Chloroquine and Hydroxychloroquine in Infectious, Immune, Neoplastic, and Neurological Diseases: A Mini-Review. Clin Drug Investig. 2018. doi: 10.1007/s40261-018-0656-y. PubMed PMID: 29737455.
13. Misztak P, Panczyszyn-Trzewik P, Sowa-Kucma M. Histone deacetylases (HDACs) as therapeutic target for depressive disorders. Pharmacol Rep. 2018;70(2):398-408. doi: 10.1016/j.pharep.2017.08.001. PubMed PMID: 29456074.
14. Rudmann DG. On-target and off-target-based toxicologic effects. Toxicol Pathol. 2013;41(2):310-4. doi: 10.1177/0192623312464311. PubMed PMID: 23085982.
15. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW, 3rd, Su AI. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 2009;10(11):R130. doi: 10.1186/gb-2009-10-11-r130. PubMed PMID: 19919682; PMCID: PMC3091323.
16. Grover M, Camilleri M. Effects on gastrointestinal functions and symptoms of serotonergic psychoactive agents used in functional gastrointestinal diseases. J Gastroenterol.

2013;48(2):177-81. doi: 10.1007/s00535-012-0726-5. PubMed PMID: 23254779; PMCID: PMC3698430.
17. Berard A, Sheehy O, Zhao JP, Vinet E, Bernatsky S, Abrahamowicz M. SSRI and SNRI use during pregnancy and the risk of persistent pulmonary hypertension of the newborn. Br J Clin Pharmacol. 2017;83(5):1126-33. doi: 10.1111/bcp.13194. PubMed PMID: 27874994; PMCID: PMC5401975.
18. O'Neil NJ, Bailey ML, Hieter P. Synthetic lethality and cancer. Nat Rev Genet. 2017;18(10):613-23. doi: 10.1038/nrg.2017.47. PubMed PMID: 28649135.
19. Nijman SM. Synthetic lethality: general principles, utility and detection using genetic screens in human cells. FEBS Lett. 2011;585(1):1-6. doi: 10.1016/j.febslet.2010.11.024. PubMed PMID: 21094158; PMCID: PMC3018572.
20. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70. PubMed PMID: 10647931.
21. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74. doi: 10.1016/j.cell.2011.02.013. PubMed PMID: 21376230.
22. Horlbeck MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, Adamson B, Chow ED, Kampmann M, Peterson TR, Nakamura K, Fischbach MA, Weissman JS, Gilbert LA. Mapping the Genetic Landscape of Human Cells. Cell. 2018. doi: 10.1016/j.cell.2018.06.010. PubMed PMID: 30033366.
23. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol. 2018;16(9):e2006643. doi: 10.1371/journal.pbio.2006643. PubMed PMID: 30226837; PMCID: PMC6143198.
24. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS, Durocher D, Angers S, Moffat J. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015;163(6):1515-26. doi: 10.1016/j.cell.2015.11.015. PubMed PMID: 26627737.
25. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van Diemen FR, Olk N, Stukalov A, Marceau C, Janssen H, Carette JE, Bennett KL, Colinge J, Superti-Furga G, Brummelkamp TR. Gene essentiality and synthetic lethality in haploid human cells. Science. 2015;350(6264):1092-6. doi: 10.1126/science.aac7557. PubMed PMID: 26472760.
26. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. Identification and characterization of essential genes in the human genome. Science. 2015;350(6264):1096-101. doi: 10.1126/science.aac7041. PubMed PMID: 26472758; PMCID: PMC4662922.
27. Apaolaza I, José-Eneriz ES, Tobalina L, Miranda E, Garate L, Agirre X, Prósper F, Planes FJ. An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. Nature Communications. 2017;8(1):459. doi: 10.1038/s41467-017-00555-y.
28. Klamt S. Generalized concept of minimal cut sets in biochemical networks. Biosystems. 2006;83(2-3):233-47. doi: 10.1016/j.biosystems.2005.04.009. PubMed PMID: 16303240.
29. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC. Defining a Cancer Dependency Map. Cell. 2017;170(3):564-76 e16. doi: 10.1016/j.cell.2017.06.010. PubMed PMID: 28753430; PMCID: PMC5667678.
30. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169(7):1177-86. doi: 10.1016/j.cell.2017.05.038. PubMed PMID: 28622505.
31. http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death.

32. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, Norris A, Sanseau P, Cavalla D, Pirmohamed M. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2018. doi: 10.1038/nrd.2018.168. PubMed PMID: 30310233.
33. Kirkland JL, Tchkonia T, Zhu Y, Niedernhofer LJ, Robbins PD. The Clinical Potential of Senolytic Drugs. J Am Geriatr Soc. 2017;65(10):2297-301. doi: 10.1111/jgs.14969. PubMed PMID: 28869295; PMCID: PMC5641223.
34. Surface LE, Park, J., Kumar, S., Burrow, D.T., Lyu, C., Li J., Song, N., Mumm, S., Haller, G., Gu, CC, Baker, J.C., Mohseni, M., Sum, M., Huskey, M., Duan, S., Bijanki, V.N., Civitelli, R., Gardner, MJ, McAndrew, C.M., Ricci, W.M., Gurnett, C.A., Diemer, K., Zhou, Y., Rajagopal, A., Bae, Y., Lee, B.E., Carette, J., Varadarajan, M., Birsoy, K., Brummelkamp, T.R., Sabatini, D.M., O'Shea, E.K., Peterson, T.R. ATRAID, a genetic factor that regulates the action of nitrogen-containing bisphosphonates on bone. bioRxiv. 2019. doi: 10.1101/338350.
35. Yu Z, Surface LE, Park CY, Horlbeck MA, Wyant GA, Abu-Remaileh M, Peterson TR, Sabatini DM, Weissman JS, O'Shea EK. Identification of a transporter complex responsible for the cytosolic entry of nitrogen-containing-bisphosphonates. Elife. 2018;7. doi: 10.7554/eLife.36620. PubMed PMID: 29745899.
36. Sabatini DM. Twenty-five years of mTOR: Uncovering the link from nutrients to growth. Proc Natl Acad Sci U S A. 2017;114(45):11818-25. doi: 10.1073/pnas.1716173114. PubMed PMID: 29078414; PMCID: PMC5692607.
37. Peterson TR, Head, R. Personal communication on single-cell RNA seq on mouse liver. 2017.
38. Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, Harris RA, O'Hare MJ, Neville AM, Simpson AJ, Strausberg RL. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. Proc Natl Acad Sci U S A. 2003;100(8):4702-5. Epub 2003/04/03. doi: 10.1073/pnas.0831040100. PubMed PMID: 12671075; PMCID: PMC153619.
39. Garcia-Ortega LF, Martinez O. How Many Genes Are Expressed in a Transcriptome? Estimation and Results for RNA-Seq. PLoS One. 2015;10(6):e0130262. Epub 2015/06/25. doi: 10.1371/journal.pone.0130262. PubMed PMID: 26107654; PMCID: PMC4479379.
40. Kveler K, Starosvetsky E, Ziv-Kenet A, Kalugny Y, Gorelik Y, Shalev-Malul G, Aizenbud-Reshef N, Dubovik T, Briller M, Campbell J, Rieckmann JC, Asbeh N, Rimar D, Meissner F, Wiser J, Shen-Orr SS. Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. Nat Biotechnol. 2018. doi: 10.1038/nbt.4152. PubMed PMID: 29912209.
41. Skinnider MA, Stacey RG, Foster LJ. Genomic data integration systematically biases interactome mapping. PLoS Comput Biol. 2018;14(10):e1006474. doi: 10.1371/journal.pcbi.1006474. PubMed PMID: 30332399; PMCID: PMC6192561.
42. Nelson MR, Johnson T, Warren L, Hughes AR, Chissoe SL, Xu CF, Waterworth DM. The genetics of drug efficacy: opportunities and challenges. Nat Rev Genet. 2016;17(4):197-206. doi: 10.1038/nrg.2016.12. PubMed PMID: 26972588.
43. Swinney DC. Phenotypic vs. target-based drug discovery for first-in-class medicines. Clin Pharmacol Ther. 2013;93(4):299-301. doi: 10.1038/clpt.2012.236. PubMed PMID: 23511784.
44. Jost M, Weissman JS. CRISPR Approaches to Small Molecule Target Identification. ACS Chem Biol. 2018;13(2):366-75. doi: 10.1021/acschembio.7b00965. PubMed PMID: 29261286; PMCID: PMC5834945.